Label-Efficient Segmentation for Diverse Scenarios

Yunzhi Zhuge

A thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
The University of Adelaide

September 1, 2023

Contents

List of Figures

List of Tables

University of Adelaide

Abstract

Label-Efficient Segmentation for Diverse Scenarios

by Yunzhi Zhuge

Segmentation, a fundamentally important task in computer vision, aims to partition an image into multiple distinct and meaningful regions or segments. In this thesis, we analyze the importance label-efficient segmentation techniques and provide a series of methods to address segmentation tasks in different scenarios.

First, we propose a Deep Reasoning Network for few-shot semantic segmentation, termed as DRNet, which is a novel approach that relies on dynamic convolutions to segment objects of new categories. Unlike previous works that directly apply convolutional layers to integrated features to predict segmentation masks, our DRNet generates learnable parameters for predicting layers based on query features, allowing for greater flexibility and adaptability.

Second, we conduct further experiments and propose mining both dynamic and regional context, termed as DRCNet, for few-shot semantic segmentation. Specifically, we introduce a Dynamic Context Module to capture spatial details in the query images, and a Regional Context Module to model the prototypes for ambiguous regions while excluding background and ambiguous objects in query images. The superior performance of our method is demonstrated on various benchmarks.

Third, we address the unsupervised video object segmentation task by learning both motion and temporal cues, in a method termed as MTNet. The proposed MTNet integrates appearance and motion information through a Bi-modal Feature Fusion Module and models the relations between adjacent frames using a Mixed Temporal Transformer. Achieving state-of-the-art results on multiple datasets while maintaining a much faster inference speed.

Finally, we propose a semi-supervised learning method for bird's-eye-view (BEV) semantic segmentation, which represents the first attempt at performing label-efficient learning in this field.Without any whistles-and-bells, our proposed BEV-S^4 can achieve results on par with fully-supervised methods while requiring significantly fewer labels. We hope that our approach could serve as a strong baseline and potentially attract more attention to learning BEV perception with fewer labels.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Yunzhi Zhuge

September 1, 2023

Acknowledgements

Firstly, I would like to express my warmest gratitude to my supervisor, Prof. Chunhua Shen, for his unwavering support and guidance throughout my PhD career. His expertise, patience, and constructive suggestions have been invaluable in shaping my research and helping me grow as a scholar. I am truly fortunate to have had the opportunity to work under his mentorship and learn from him. His dedication to academic excellence has inspired me to strive for the highest standards in my work.

Secondly, I would also like to extend my heartfelt appreciation to my colleagues, who have provided a stimulating and supportive environment during my research journey. They are Zhi Tian, Xinlong Wang, Tong He, Chen Hao, Jia-wang Bian, Hu Wang, Wei Yin, Bowen Zhang, Bohan Zhuang, Xinyu Wang, Yifan Liu and Yutong Dai. Their encouragement and insightful discussions have enriched my academic experience and played a crucial role in the successful completion of my thesis.

Then, I cannot express enough gratitude to my parents who have given me their constant love, encouragement, and unwavering belief in my abilities. Their sacrifices and support have given me endless power to pursue my PhD degree with determination and resilience.

Lastly, I would like to convey my heartfelt gratitude to everyone who has provided support and guidance during my PhD journey. Their contributions have been indispensable in making this achievement possible.

Publications

This thesis contains the following works that have been published or prepared for publication:

- Deep Reasoning Network for Few-shot Semantic Segmentation.
  Yunzhi Zhuge, Chunhua Shen.
  ACM International Conference on Multimedia, 2021.
- Few-shot Semantic Segmentation by Exploiting Dynamic and Regional Contexts.
  Hongyu Gu*, **Yunzhi Zhuge***, Lu Zhang, Jinqing Qi, Huchuan Lu.
  International Conference on Multimedia & Expo, 2023.
- Learning Motion and Temporal Cues for Unsupervised Video Object Segmentation.
  Yunzhi Zhuge, Hongyu Gu, Lu Zhang, Jinqing Qi, Huchuan Lu.
  Under review of ACM International Conference on Multimedia, 2023.
- BEV-S^4: Semi-supervised Bird's-eye-view Semantic Segmentation.
  Yunzhi Zhuge, Jiayuan Zhou, Lijun Wang, Yifan Wang, Huchuan Lu.
  Under review of ACM International Conference on Multimedia, 2023.

In addition, I have the following papers not included in this thesis:

- Multi-granularity Transformer for Image Super-resolution.
  Yunzhi Zhuge, Xu Jia.
  Asian Conference on Computer Vision, 2022.

Chapter 1

Introduction

Dense prediction tasks in computer vision, such as object detection and image segmentation, are fundamentally important as they form the basis for interpreting and understanding environments. In the era of deep learning, and particularly with the recent explosion of large foundation models, we are thrilled to witness many tasks achieving astonishing results that were previously difficult to imagine. Typically, the design of prediction tasks consists of two aspects: model engineering and data engineering. Model engineering involves designing more effective neural networks to improve the performance and efficiency of prediction tasks. On the other hand, data engineering focuses on collecting and annotating diverse datasets that can address existing challenges or help solve new tasks in computer vision. Given that preparing new datasets is labor-intensive and challenging, many researchers opt for the first route and concentrate their efforts on designing more effective or efficient models.

Some works also aim to reduce the reliance on labels by exploring label-efficient learning paradigms, such as few-shot, unsupervised, self-supervised, weakly-supervised, and semi-supervised learning. These learning paradigms focus on situations with limited labeled data, making the most of the available information to learn accurate models. Although there are differences in the way they handle data availability, problem settings, and specific goals, these learning paradigms are all important as they address the challenges associated with data availability, labeling effort, and generalization in perception tasks.

On the other hand, visual perception plays a vital role in autonomous driving, particularly in the pure vision-based technical route, which has already been broadly employed in advanced driver-assistance systems (ADAS) such as Tesla's Autopilot (Wikipedia contributors, 2023). Currently, many researchers focus on the bird's-eye view (BEV) coordinate system, which provides an overhead view of the surrounding environment from a bird's-eye perspective. These approaches often rely on large-scale datasets with precise annotations, which can be labor-intensive and expensive to obtain. Therefore, finding ways to develop label-efficient perception models is of great importance By leveraging label-efficient learning paradigms, more cost-effective and scalable perception systems in the context of BEV-based autonomous driving can be achieved.

In this thesis, our goal is to design label-efficient paradigms for both conventional 2D images and autonomous driving scenarios, such as few-shot segmentation and semi-supervised learning for BEV semantic segmentation. We capture relationships between support and query sets using flexible dynamic convolutions, which is a pioneering effort in this field. Furthermore, we develop a semi-supervised bird's-eye-view semantic segmentation paradigm to investigate solutions for autonomous driving that emphasize label efficiency. Additionally, we introduce a novel unsupervised video object segmentation method that leverages both motion and temporal cues to enhance segmentation performance. Through these contributions, we aim to advance the state-of-the-art in label-efficient learning, ultimately facilitating more cost-effective and scalable perception systems in various computer vision applications.

1.1 Contribution and Outline

With the objective of general scenarios and label-efficient segmentations, we design a serious of algorithms towards solving challenges in few-shot semantic segmentation, unsupervised video object segmentation and semi-supervised bird's-eye-view semantic segmentation. The main contributions of this thesis are as follows:

- The first attempt to apply dynamic convolution to solve few-shot semantic segmentation, termed as DRNet. Different from previous methods, the learnable parameters of our proposed predicting layer are dynamically generated based on support features, thereby adaptively bridging the gap between the query set and support set for a more comprehensive understanding of semantics in foreground regions. By leveraging dynamic convolutions in this manner, the model can more effectively capture the relationship between the support and query images, leading to improved performance in few-shot semantic segmentation.

- A further investigation of both dynamic and regional contexts for few-shot semantic segmentation. Specifically, we propose a dynamic context module and a regional context module. The dynamic context module is responsible for extracting spatial information from query features, while the regional context module is designed to address ambiguous regions in query images. By incorporating both modules into our model, we aim to produce more reliable results in few-shot semantic segmentation tasks.

- We address the challenge of unsupervised video object segmentation by simultaneously exploiting motion and temporal cues. To achieve this, we carefully design a bi-modal fusion model and a mixed temporal transformer, which allows our algorithm to effectively combine the strengths of both motion and temporal information.Our proposed method achieves new state-of-the-art results on benchmark datasets for unsupervised video object segmentation. Moreover, it maintains real-time inference speed on a 2080ti GPU

- An empirical investigation of solving bird's-eye-view (BEV) semantic segmentation with semi-supervised learning. For the first time, we propose a learning paradigm for BEV perception tasks under label scarce situations.We are excited to discover that our semi-supervised learning method can achieve on par results with the fully supervised method while using only 10% of labeled data.

Chapter 2 provides the necessary background information for understanding the problems and techniques addressed in this thesis, which encompasses a range of topics such as few-shot semantic segmentation, unsupervised video object segmentation, and bird's-eye-view semantic segmentation. In this chapter, we review related tasks and their respective challenges, including semantic segmentation, few-shot semantic segmentation, semi-supervised semantic segmentation, video object segmentation and bird's-eye-view perception. By reviewing these related tasks and the techniques used to address them, this chapter aims to provide the necessary context for understanding the novel contributions and findings presented in the subsequent chapters of the thesis.

Chapter 3 highlights the shortcomings of previous prototype-based few-shot semantic segmentation algorithms. Specifically, the relation modeling between the support set and query set is relatively fixed, which may struggle to handle inter-class gaps between training and testing categories. To address this issue, we introduce dynamic convolutions via the DRNet (Zhuge and Shen, 2021), which has been shown to yield more robust results.

In **Chapter** 4, we addresses the limitations associated with masked average pooling as well as the challenges posed by noisy backgrounds and ambiguous regions in query images. To tackle these issues, we introduce the dynamic context module and the regional context module. These novel approaches lead to competitive results, both quantitatively and qualitatively, as evidenced by the performance metrics and visual analysis presented in (Gu et al., 2023).

In **Chapter** 5, we introduces a cutting-edge approach to address the challenges of unsupervised video object segmentation. Our method capitalizes on motion cues through compact bi-modal fusion modules and exploits temporal cues via mixed temporal transformer modules. This innovative strategy not only achieves state-of-the-art results, but also operates at an impressive speed of approximately 45 fps on a 2080ti GPU. Remarkably, our method is three times faster than the previously best-performing technique (Zhuge et al., 2023b).

In **Chapter** 6, we focus on bird's-eye-view (BEV) semantic segmentation, a crucial task within the realm of BEV perception. Our approach employs a label-efficient paradigm to tackle this problem. As a baseline, we utilize the fully-supervised, yet straightforward and effective, PETR-v2 method (Liu et al., 2022b). Under this novel setting, we explore various semi-supervised learning techniques, such as teacher-student networks and data augmentation strategies. Remarkably, even without incorporating complex components, our approach achieves results comparable to those of supervised counterparts (Zhuge et al., 2023a).

In **Chapter** 7, we provide a summary of the thesis, highlighting the key findings and contributions made throughout the work. Furthermore, we discuss potential avenues for future research that is related to the foundation laid by our study.

Chapter 2

Background

2.1 Semantic Segmentation

Semantic segmentation (Guo et al., 2018; Hao, Zhou, and Guo, 2020) is a crucial task in the field of computer vision that aims to allocate semantic labels to each pixel within an image, thereby enabling the categorizing of each pixel into a predefined category of objects. With the rise of deep learning techniques, the Fully Convolutional Network (FCN) architecture (Long, Shelhamer, and Darrell, 2015) has emerged as the predominant approach for semantic segmentation, utilizing a holistic and efficient strategy for pixel-to-pixel classification. Since then, various works emerge to solve the semantic segmentation problem based on FCN. Some works focus on enlarging the receptive fields such as dilated convolutions (Yu and Koltun, 2015; Chen et al., 2017a; Mehta et al., 2018), pyramid pooling (Chen et al., 2017b; Zhao et al., 2017; Hou et al., 2020) and non-local operations (Huang et al., 2019; Zhu et al., 2019; Yu et al., 2020), which enabled the model to have a more comprehensive understanding of the image.

Motivated by the prevalence and success of Vision Transformers (Carion et al., 2020; Dosovitskiy et al., 2020), numerous recent studies have investigated the integration of transformers for addressing the task of semantic segmentation. SETR (Zheng et al., 2021) first leveraged the capabilities of transformers to solve semantic segmentation by approaching semantic segmentation as a sequence-to-sequence problem, showing promises of research. SegFormer (Xie et al., 2021a) made a further step by adopting hierarchical transformer architecture and designing more compact decoder, thereby achieving superior results with significantly reduced computational expenses. HRViT (Gu et al., 2022) is a vision transformer backbone specifically optimized for semantic segmentation. By combining the merits of ViTs and HRNet, the model is capable of learning semantically-rich and spatially-precise multi-scale representations in an efficient manner. MaskFormer (Cheng, Schwing, and Kirillov, 2021) adopted the set prediction mechanism proposed in DETR (Carion et al., 2020) and replaced the traditional per-pixel classification model with a mask classification model, enabling it to effectively address both semantic-level and instance-level segmentation challenges in a seamless pipeline. However, both the aforementioned methods necessitate substantial amounts of annotated data to attain high performances, which is time-consuming, labor-intensive, and expensive. Furthermore, those fully-supervised models may lack generalizability to new or unseen data. The most recent study SAM (Kirillov et al., 2023) leveraged an extensive dataset comprising millions of images and billions of annotated masks. This comprehensive approach not only yields impressive performance in various image segmentation tasks but also exhibits robust zero-shot transfer capabilities.

2.2 Few-shot Semantic Segmentation

The objective of few-shot semantic segmentation (FSS) is to obtain the ability of performing semantic segmentation on a query image utilizing a limited number of support images that have been labeled through pixel-level annotation. (Shaban et al., 2017) first raised the problem.The paradigm is comprised of a conditional branch and a segmentation branch, where the conditional branch generates classifier weights to be utilized for segmenting the query image. Majority of the following methods inherit the dual-branch architecture and are based on prototypical learning (Snell, Swersky, and Zemel, 2017), where the representative works include SG-One (Zhang et al., 2020a), CANet (Zhang et al., 2019b), PGNet (Zhang et al., 2019a), PANet (Wang et al., 2019a), PFENet (Tian et al., 2020), PPN (Yang et al., 2020a) and PMM (Liu et al., 2020c). The fundamental dissimilarities between these methods lie in the means for acquiring and utilizing the prototypes. CANet concatenated the support prototype with the query features and implemented an iterative optimization module to progressively refine the prediction. PFENet introduced the concept of utilizing high-level categorical information and employed feature pyramid fusion to boost the fused features. PPN decomposed the comprehensive class representation into a collection of part-aware prototypes to capture the fine-grained and varied object feature.

Other lines of research either focuses on capturing fine-gained correspondence relations or calculating cross-attention between the query and the support images. FSNet (Min, Kang, and Cho, 2021) assembled a collection of 4D correlation tensors by utilizing a wide range of geometric and semantic feature representations extracted from multiple intermediate layers of a convolutional neural network, thereby providing a comprehensive set of correspondences across various visual aspects. In (Hong et al., 2022), a volumetric transformer module for the cost aggregation incorporating a 4D swin transformer was proposed to effectively capture the hypercorrelation in a volumetric context. CyCTR (Zhang et al., 2021a) extracted information from the support image by limiting the attention of the query features solely to cycle-consistent support features, thereby reducing the impact of noise. CATrans (Zhang et al., 2022) developed two types of transformer blocks, named relation-guided context transformer and relation-guided affinity transformer to transfer informative semantic information from support to query image and accurately determine the cross-correspondences respectively. In this study, we probe the feasibility of using dynamic convolutional layers to capture the relationships of support and query images for few-shot semantic segmentation.

2.3 Semi-supervised Semantic Segmentation

Semi-supervised semantic segmentation represents a learning approach in which the objective is to derive semantic segmentation models through the utilization of a limited subset of labeled data, with the majority of the data remaining unlabeled. The notion of consistency regularization has received considerable attention in the field of semi-supervised semantic segmentation, the essence of which is to ensure the coherence of predictions and intermediate features in the presence of multiple perturbations. (Kim et al., 2020) considered the inter-pixel correlation and proposed a structured consistency loss which enabled the network to learn more powerful generalization capabilities to predict in harmony with neighboring pixels. PseudoSeg (Zou et al., 2020) introduced a novel formulation of pseudo-labeling, where it derives structured pseudo-labels for supplementary data, illustrating that the utilization of well-calibrated soft pseudo-labels obtained can significantly enhance the effectiveness of consistency training in semantic segmentation. ClassMix (Olsson et al., 2021) proposed a unique data augmentation strategy for semantic segmentation, which involves the application of a cut-and-paste technique to selectively transfer half of the predicted classes from one image to another. CCT (Ouali, Hudelot, and Tami, 2020) enforced consistency in the predictions of the main decoder on unlabeled data and those of the auxiliary decoders, with the aim of improving the representation learning capabilities of the main decoder. Similarly, in CPS (Chen et al., 2021), the consistency constraint was imposed on two segmentation networks, where the pseudo segmentation map generated by one network serves as an additional supervision signal for the other. DARS (He, Yang, and Qi, 2021) dedicated to address the issue of bias in pseudo-labels by proposing a technique that combines distribution alignment and random sampling to re-balance the skewed pseudo-labels and harmonize their distribution with the actual distribution. U^2-PL (Wang et al., 2022) utilized both reliable and unreliable pixels in a comprehensive manner by using reliable predictions to derive positive pseudo-labels and treating unreliable pixels as negative samples. (Liu et al., 2022d) observed that consistency learning methods vulnerable to the influence of inaccurate predictions of unlabelled training images. To address this issue, they introduced a novel mean-teaching framework that employs an auxiliary teacher and a confidence-weighted cross-entropy loss, aimed at enhancing the generalization of consistency learning.

2.4 Video Object Segmentation

Video object segmentation, which involves identifying and segmenting objects with specific properties within a video scene, plays a crucial role in a broad spectrum of downstream applications such as autonomous driving, robotics, virtual reality, video rendering, and online meetings. To date, heuristic knowledge-based methods (Papazoglou and Ferrari, 2013; Wang et al., 2017) and hand-crafted feature-based methods have become outdated. Consequently, our focus is primarily on deep learning-based approaches, owing to their remarkable performance and widespread adoption in the field. In the deep learning era, video object segmentation can generally be classified into four categories based on the level of human interaction during inference. These categories include unsupervised video object segmentation(UVOS), semi-supervised video object segmentation(SVOS), referring video object segmentation(RVOS), and interactive video object segmentation(IVOS), each differing in their degree of human involvement and the supervisions provided for the task. Semi-supervised methods for video object segmentation primarily rely on either mask propagation approaches (Perazzi et al., 2017; Jang and Kim, 2017; Jampani, Gadde, and Gehler, 2017; Khoreva et al., 2019; Xiao et al., 2018) or matching-based solutions (Oh et al., 2019b; Zhang et al., 2020b; Yang, Wei, and Yang, 2020; Seong, Hyun, and Kim, 2020; Xie et al., 2021d; Cheng, Tai, and Tang, 2021b). The former utilized previous frame masks to infer the current mask, a process which is prone to error accumulation due to occlusions and drifts during mask propagation. On the other hand, matching-based methods focused on constructing an embedding space that captures the object embeddings of initial and previous masks. These methods determined the label of each pixel by assessing its similarity to the target object within the embedding space, thereby mitigating some of the issues faced by mask propagation techniques. Referring video object segmentation (Gavrilyuk et al., 2018; Khoreva, Rohrbach, and Schiele, 2019; Seo, Lee, and Han, 2020) is an emerging topic in which the object to be segmented is defined by a linguistic sentence. Inspired by the success of vision transformers (Carion et al., 2020; Dosovitskiy et al., 2020), numerous recent works (Botach, Zheltonozhskii, and Baskin, 2022; Wu et al., 2022; Liang et al., 2023) were proposed to capture visual and textual context using self-attention and cross-attention mechanisms. Interactive video object segmentation (Oh et al., 2019a; Heo, Koh, and Kim, 2021; Cheng, Tai, and Tang, 2021a; Yin et al., 2021) aims to aid the model in refining prediction results by incorporating human-in-the-loop through multiple rounds of prediction and refinement.

In contrast to these approaches, the goal of unsupervised video object segmentation is to automatically segment objects in a video solely based on the visual content, without any human intervention or prior knowledge about the objects in the scene. Since the background and various objects can be ambiguous, unsupervised video object segmentation is considered more challenging, and thus, requires greater effort. In the realm of deep learning, particularly following the fully convolutional network (FCN) (Long, Shelhamer, and Darrell, 2015), which addresses semantic segmentation through per-pixel predictions, UVOS (Tokmakov, Alahari, and Schmid, 2017) has experienced substantial advancements and progress. Drawing inspiration from non-local networks (Wang et al., 2018b), several approaches, including ADNet (Yang et al., 2019), COSNet (Lu et al., 2019), AGNN (Wang et al., 2019b) and F2Net (Liu et al., 2021a), modeled inter-frame correspondence to extract global information and achieve a more comprehensive understanding of video content. On the other hand, optical flow supplies vital motion information for the localization and differentiation of primary objects, serving as an additional cue in UVOS. MATNet (Zhou et al., 2020b) presented a two-stream interleaved encoder, providing a motion-to-appearance pathway for information propagation and a Motion-Attentive Transition Module for feature selection. AMC-Net (Yang et al., 2021b) proposed a co-attention gate for motion appearance re-weighting and an adaptive motion correction module for feature fusion. Employing a full-duplex strategy, FSNet (Ji et al., 2021) designed a relational cross-attention module and a bidirectional purification module to effectively fuse appearance and motion information. DTNet (Zhang et al., 2021b) implemented an optimal structural matching approach for the purification and alignment of motion-appearance features. HFAN (Pei et al., 2022) introduced a hierarchical feature alignment network that aligns appearance-motion features with primary objects and adaptively fuses them to enhance performance. Although these two-stream methods demonstrate satisfactory performance in certain scenarios, they struggle to track primary objects without temporal contexts, particularly in intricate occlusion scenes.

2.5 Camera-based BEV Perception

Camera-based BEV perception has gained significant attention in both industry and academia due to its unique advantages in assisting autonomous driving in a sensor-friendly manner. Representing surrounding objects and environments in BEV is beneficial for subsequent tasks, such as planning and control. Compared to LiDAR-based or fusion-based solutions, camera-only BEV perception has attracted more interest from researchers owing to its uncurated characteristics. However, the absence of accurate depth information presents challenges.

View transformation, a vital component in vision-based BEV perception, plays a key role in constructing 3D information and encoding 3D priors from 2D features. Camera-based BEV perception encompasses several aspects, including 3D object detection, BEV semantic segmentation, and motion prediction. Among these, BEV semantic segmentation aims to assign each pixel in the BEV map with a semantic label, making it crucial for various applications that necessitate understanding complex environments from an overhead perspective.

VPN (Pan et al., 2020) represents a groundbreaking contribution to this field. It was the first to leverage simulation environments to gather cross-view annotations. Since then, numerous methods (Roddick and Cipolla, 2020; Philion and Fidler, 2020; Ng et al., 2020; Zhou and Krähenbühl, 2022; Xie et al., 2022) have emerged, consistently setting new performance benchmarks. However, the high cost of annotating data may hinder the development of larger datasets, potentially impeding progress in this field.

Statement of Authorship

Title of Paper	Deep Reasoning Network for Few-shot Semantic Segmentation
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Published by ACM MM 2021

Principal Author

Name of Principal Author (Candidate)	Yunzhi Zhuge		
Contribution to the Paper	Proposed the ideas, analysed the feasibility, conducted experiements and wrote the manuscript of the paper.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	2023/4/26

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen		
Contribution to the Paper	Discussion, Revision of the paper		
Signature		Date	2023/4/27

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Please cut and paste additional co-author panels here as required.

Chapter 3

Deep Reasoning Network for Few-shot Semantic Segmentation

3.1 Introduction

In dense prediction tasks such as semantic segmentation, instance segmentation and video object segmentation, collecting labeled data is often a tedious and expensive process. Furthermore, the labeled data is often limited to fixed set of predefined categories, which make it difficult for the model to generalize to unseen categories.

To address these challenges, few-shot learning has been considered as a promising solution. Specifically, models trained on previous tasks are expected to generalize to unseen tasks given only several labeled images as prompt. This allows the transfer of knowledge across tasks and facilitates more efficient use of the labeled data. Few-shot learning has shown great potential in overcoming the limitations of previous deep learning approaches, particularly in scenarios where labeled data is scarce or costly to obtain.

Few-shot semantic segmentation (FSS) is situated at the intersection of few-shot learning and semantic segmentation, with the goal of accurately segmenting the foreground regions of a novel object category using limited training data, typically comprising only a few image-mask pairs. One of the key objectives of FSS is to effectively leverage the information associated with the foreground objects, while mitigating the impact of the background regions that can potentially hinder the segmentation process. To accomplish this, FSS relies on the transfer of information from support images to query images, with the aim of accurately segmenting objects that belong to the same categories across both image sets. However, due to the scarcity of available annotated data, achieving this task requires a sophisticated and robust approach that can effectively overcome these limitations.

In this chapter, we present a novel approach to solve few-shot semantic segmentation (FSS) by adopting dynamic convolutions to model the relationship between query and support images. This technique employs a set of learnable filter coefficients to generate a set of weights for a fixed set of filters, which are subsequently applied to

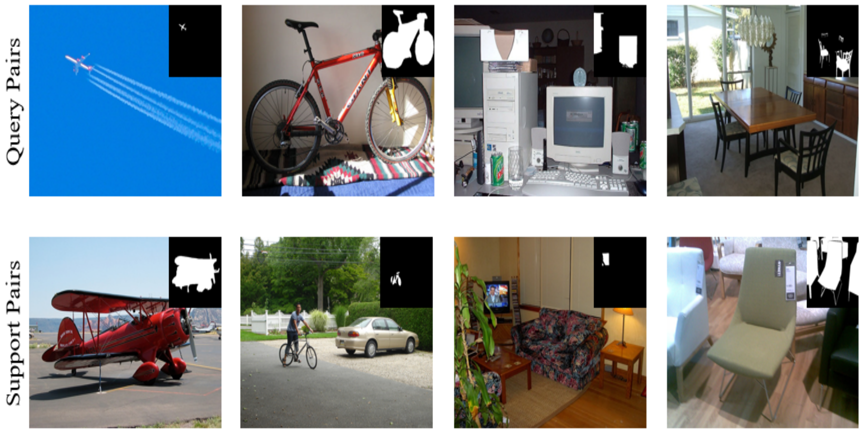


Figure 3.1. **Examples of support image images and query image pairs.** There are many challenges, such as complex relations, various appearances, scale changes.

distinct regions of the input. Specifically, our research introduces a novel approach, the Deep Reasoning Network (DRNet), which is designed to generate the parameters of predicting layers and accurately infer the segmentation mask for every unseen category in an adaptive manner. By this way, our method can effectively capture the critical interactions between support and query images, which is a crucial aspect of accurate FSS. Specifically, the dynamic convolutions enable the model to learn the optimal filter coefficients for each query image, thus ensuring that the weight generation process is customized to each specific input.

More concretely, we introduce an Attentional Feature Integration Sub-network (AFIS) to extract consistent features from both support and query images. This approach serves as an essential component of our methodology, with shared weights facilitating category coherence across different data streams. Then a Pooling-based Guidance Module (PGM) is employed to progressively establish correlations between support features and query features. This is achieved through a progressive process, which enables the identification and strengthening of correlations between the aforementioned features. To disseminate information from support images to various query images, we further propose a Dynamic Prediction Module (DPM) for generating the parameters of predicting layers. The proposed modules are unified for the deep reasoning of each query image segmentation.

In order to facilitate the dissemination of information from support images to various query images, we introduce a novel approach called the Dynamic Prediction Module (DPM). This module is designed to generate parameters for predicting layers, which are used to inform the segmentation of each query image. Our proposed method is unified in its application, enabling deep reasoning for the segmentation of all query images. Experiments on two public benchmarks have demonstrated that our approach achieves superior performance and outperforms the very recent state-of-the-art methods.

3.2 Background

Semantic segmentation (Chen et al., 2017a; Xie et al., 2021a) is essentially important in the field of computer vision, where the objective is to classify each pixel in an image into predefined categories. This task is critical for a wide range of applications, including autonomous driving (Hu et al., 2023), medical imaging (Singh, Sengupta, and Lakshminarayanan, 2020), robot navigation (Gupta et al., 2020) and image manipulation (Zhou et al., 2020a). However, in the era of deep learning, the performance of semantic segmentation models is often highly dependent on the scale and quality of the training dataset. Unfortunately, generating these datasets is a laborious and time-consuming process, as it requires annotating images at the pixel level. This annotation process is not only expensive in terms of human resources but also prone to errors due to the subjective nature of the task. Therefore, it is important to find ways that can decrease the dependence on human-labeled data for semantic segmentation to make it more practical and scalable.

To reduce the reliance on human annotations, a wide range of techniques have been explored, such as unsupervised or weakly supervised learning, to learn representations from unlabeled or partially labeled data. While these methods have shown to be promising, their performance still significantly lags behind that of fully supervised methods in terms of accuracy and robustness.

In addition to the reliance on human annotations, semantic segmentation still struggles with the generalization problem. The models trained on datasets with specific categories exhibit limited adaptability to novel object classes and scenes. The above problem can pose a severe challenge for its applicability to various downstream applications, such as autonomous driving, where the scenes are continually evolving, and the availability of labeled data for new environments or objects may be limited.

Few-shot Semantic Segmentation (FSS) is a promising direction for addressing the aforementioned limitations in semantic segmentation, which involves the task of accurately segmenting objects of previously unseen categories with minimal annotated data, typically consisting of only one or a few samples per class. This presents a significant challenge that necessitates innovative approaches to overcome the limitations of insufficient labeled data and achieve high-quality segmentation outcomes. Specifically, FSS delivers several benefits over traditional semantic segmentation paradigms. To begin with, FSS significantly reduces the requirement of large-scale annotated data, thus overcoming the scalability and labeling cost in traditional semantic segmentation. Another significant advantages of FSS is its capability in handling the generalization problem by learning to adapt to novel categories prompted by a small set of labeled examples. This enables the model to learn more effectively from a smaller dataset and enhances its ability to generalize to unseen data, which is crucial in many real-world applications.

However, the problem of Few-shot Semantic Segmentation (FSS) is especially difficult due to the nature of the test data, which consists of novel categories that were excluded in the training set. Moreover, the huge variations in appearance and shape between the support and query images further aggravate the challenge. Thus, accurately segmenting objects across such diverse categories requires innovative solutions that can effectively transfer information from support to query images despite the variations in their features.

One main aim in Few-shot Semantic Segmentation (FSS) is to make full use of information associated with the foreground objects, while simultaneously suppressing the influence of the background regions that could impede the segmentation process. In essence, FSS requires the transfer of information from support images to query images, with the aim of accurately segmenting objects that belong to the same categories across both sets. Achieving this requires a sophisticated procedure that can overcome the limitations imposed by the scarcity of available annotated data.

Technically, FSS performs semantic segmentation on unseen object categories with only several pixel-level annotated pairs (Dong and Xing, 2018; Zhang et al., 2019b; Li et al., 2020). To solve this task, existing methods are mainly based on prototypical learning where a dual branch architecture is employed to process support images and query images. Specifically, the support branch is used to extract class prototypical information and guide the query branch for segmenting query images. To achieve this goal, Global Average Pooling (GAP) is commonly adopted to generate support vectors (Zhang et al., 2019b; Zhang et al., 2019a; Siam, Oreshkin, and Jagersand, 2019). These methods have already shown expressive results. However, there still exist several key problems in existing methods. Firstly, those prototype-based methods (Zhang et al., 2019b; Zhang et al., 2019a; Zhang et al., 2020a; Siam, Oreshkin, and Jagersand, 2019) simply integrate support features and query features after the feature extraction which neglect the internal relations between features. This is not sufficient to locate the areas that contain objects in query images. Furthermore, there exist inter-class gaps between training categories and testing categories. To solve this problem, Yang (Yang et al., 2020b) propose a online refinement strategy to adapt the network to unseen categories. However, with their model, the computational cost increases significantly while the performance is not enhanced remarkably.

To resolve the aforementioned limitations, in this work we propose a Deep Reasoning Network (DMNet) for effective FSS. First, we propose an Attentional Feature Integration Sub-network (AFIS) extract multi-level consistent features from support images and query images. With shared weights of two branches, it explores and captures more correlated information, stimulating the category consistency of different data streams. Besides, we propose a Pooling-based Guidance Module (PGM) to further enhance the relation of support set and query set in semantic level. The PGM could generate and merge features with different resolutions in a progressive manner. Finally, we propose a Dynamic Predicting Module (DPM) to reduce the gaps between training categories

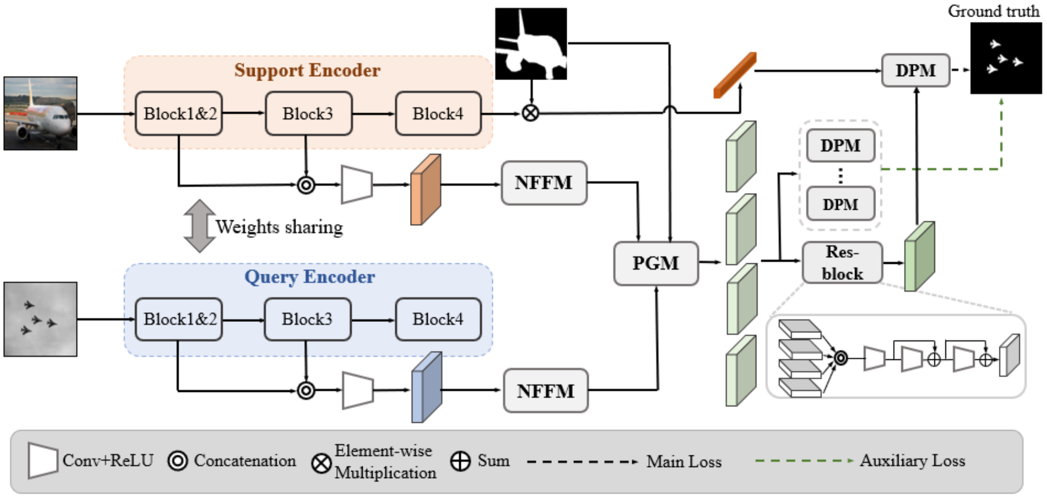


FIGURE 3.2. **Pipeline of the proposed Deep Reasoning Network (DRNet).** Our method mainly consists of Non-local Feature Fusion Module (NFFM), Pooling-based Guidance Module (PGM) and Dynamic Predicting Module (DPM). The NFFM firstly integrates and refines multi-level features in each branch. Then, the PGM exploits guidance information from support branch and collaborate with query features in different levels. Finally, the DPM conditionally generates learning parameters of predicting layer for segmentation.

and testing categories. The DPM adaptively generates the parameters of predicting layer for segmenting each query image. To verify the effectiveness of our proposed method, we conduct extensive experiments on two benchmark datasets (PASCAL-5^i (Shaban et al., 2017) and MS COCO-20^i (Nguyen and Todorovic, 2019)). The proposed method outperforms other state-of-the-art methods by a large margin. Furthermore, extensive ablation studies are implemented to demonstrate the contribution of each component module in our work.

Our main contributions can be summarized as follows:

- We propose a Deep Reasoning Network, to solve challenging problems in FSS. Different from previous methods, the prediction layers of our method are dynamically generated for each query image.
- We propose a novel pooling-based guidance module to incorporate multi-level support information into query features. This module significantly helps to precisely locate the query objects.
- Our method achieves state-of-the-art results on two public benchmarks, i.e., PASCAL-5^i and MS COCO-20^i datasets. Ablation experiments also demonstrate the effectiveness of each module in our work.

3.3 Our Approach

As shown in Figure 3.2, the proposed DRNet is composed of two branches, i.e., the support branch and query branch. Previous methods directly apply convolutional layers on the integrated features to predict the segmentation masks (Zhang et al., 2019b; Liu et al., 2020b; Wang et al., 2020). In contrast to existing approaches, our proposed Deep Reasoning Network (DRNet) offers a novel and adaptive solution for learning and predicting segmentation masks for previously unseen object categories. Specifically, DRNet generates learnable parameters of predicting layers based on the query features, allowing for greater flexibility and adaptability. Our method employs a weight-shared encoder to extract support and query features, which are then processed by the Non-Local Feature Fusion Module (NFFM) to refine the features of each branch. Additionally, our Progressive Guidance Module (PGM) integrates query and support features at different scales to more fully leverage guidance information from the query features. Finally, DRNet employs the Dynamic Prediction Module (DPM) to generate the predicting filters conditioned on the support features. The segmentation mask of each query image is obtained by the predicting filter

3.3.1 Attentional Feature Integration Sub-network

As is demonstrated in (Zhang et al., 2019b), the utilization of features from higher layers that contain more object-level concepts can result in a reduction in performance for semantic segmentation. To address this issue, we propose a multi-level feature integration method to enhance feature representation capabilities, as shown in Fig 3.2. Meanwhile, the recent proposed non-local module (Wang et al., 2018b) could capture long-range dependencies in an image or video. It also can be treated as a feature fusion module. However, the vanilla non-local module holds high complexity of matrix multiplications. To reduce the complexity, we introduce a new Non-local Feature Fusion Module (NFFM) to enhance features in each branch.

Standard Non-local Module. In a standard non-local module, three 1×1 convolutions $Conv_\phi$, $Conv_\theta$, and $Conv_\gamma$ transform input feature $X \in \mathcal{R}^{C \times H \times W}$ to new embeddings $\phi \in \mathcal{R}^{\hat{C} \times H \times W}$, $\theta \in \mathcal{R}^{\hat{C} \times H \times W}$ and $\gamma \in \mathcal{R}^{\hat{C} \times H \times W}$ as

$$\phi = Conv_\phi(X), \theta = Conv_\theta(X), \gamma = Conv_\gamma(X) \tag{3.1}$$

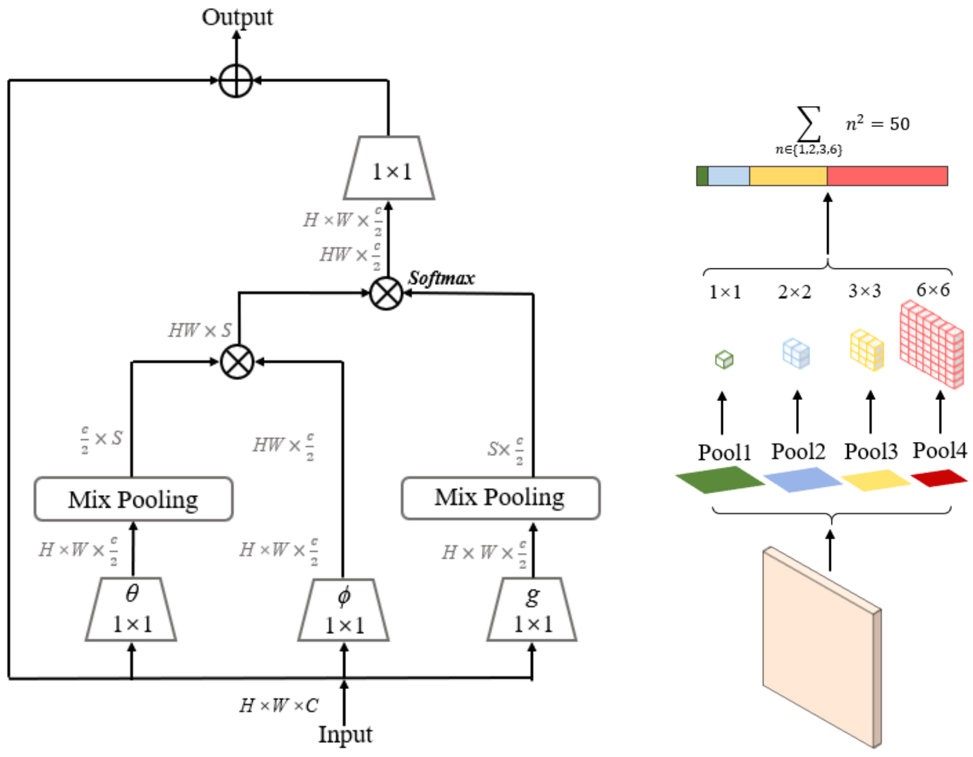
Three embeddings are then flattened to size $\hat{C} \times N$ ($N = H\dot{W}$). The unified similarity matrix V can be obtained by

$$V = Softmax(\phi^T \times \theta) \tag{3.2}$$

where $Softmax$ is the Softmax normalization. Thus, the output of the non-local module is

$$Y = Conv((V \times \gamma^T)^T) + X \tag{3.3}$$

where $V \times \gamma^T$ is to calculate the attention weight for each location in γ and $Conv$ is a 1×1 convolution to ensure the same size of the output features.



(a) (b)

FIGURE 3.3. **The proposed Non-local Feature Fusion Module (NFFM).** (a) is the overall architecture and (b) shows the detailed structure of pyramid pooling.

Reducing Complexity with Pyramid Pooling. Technically, the standard non-local module introduces high computational cost with the two matrix multiplications, resulting in $\mathbb{C}(CN^2)$=$\mathbb{C}(CW^2H^2)$ complexity. Following (Zhu et al., 2019), we reduce complexity by sampling representative points from θ and γ via Spatial Pyramid Pooling (SPP) (Lazebnik, Schmid, and Ponce, 2006). The proposed method is efficient and able to represent multi-scale relations. As shown in Figure 3.3, SPP is applied on the embeddings θ and γ to extract compact samples. We set the output size of pyramid pooling to $n \subseteq \{1, 2, 3, 6\}$, and thus the number of output samples is

$$S = \sum_{n\in\{1,2,3,6\}} n^2 = 50. \tag{3.4}$$

Considering that the spatial locations in the input features is $N = 64 \times 64 = 4096$, the complexity of matrix multiplication can be reduced by $T = \frac{N}{S} \approx 81$ times.

3.3.2 Pooling-based Guidance Module

To extract category information from support samples, previous methods usually apply a Global Average Pooling (GAP) on the support features to obtain an embedding vector. However, directly correlating the embedding vector with query features is sub-optimal due to the mismatching of support masks and query objects. To address

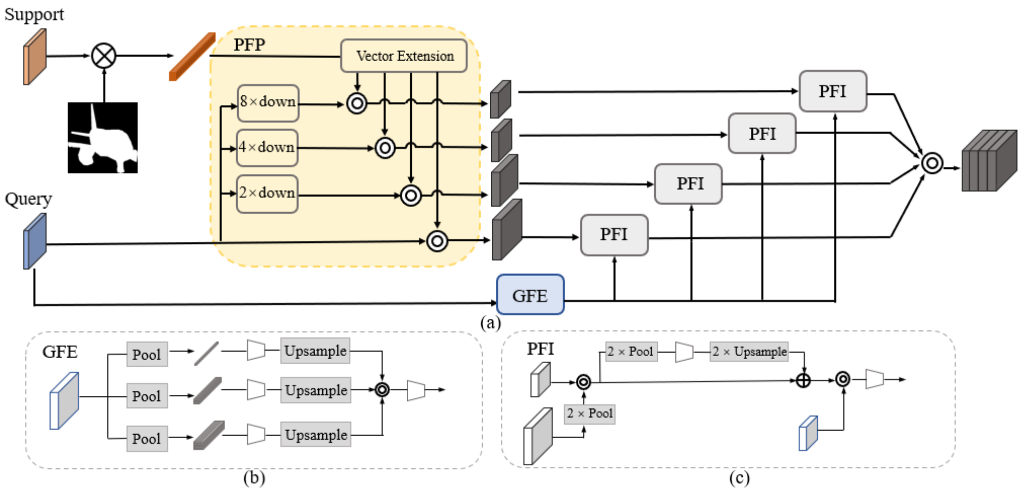


FIGURE 3.4. **The Pooling-based Guidance Module.** (a) is the whole architecture. (b) and (c) are details of Global Feature Extraction and Progressive Feature Integration

this problem, we propose a Pooling-based Guidance Module (PGM) to enhance the spatial consistency and exploit the guidance of support samples. More specifically, the PGM takes the query features, support features and support masks as inputs. By using GAP, embedding vectors correspond to category information are extracted from the foreground area of support features. Those vectors are further expanded and integrated with query features of different resolutions in a progressive manner.

As shown in Figure 3.4, there are three key components for PGM: 1) Pyramid Feature Fusion (PFF) first down-samples query features into different scales, and then integrates with category-oriented features in each scale; 2) Global Feature Extraction (GFE) is performed to extract global guidance information from fused pyramid features; 3) Progressive Feature Integration (PFI) is used to merge multi-level features.

As is shown in Figure 3.1, there exists huge variances between support masks and query masks. Figure 3.5 statically analyzes the variance ratios in each split. It shows that most support-query pairs are inconsistent in their spatial size. Directly integrating support features with query features might lead to poor localization results. Thus, we propose the following pyramid feature fusion.

Pyramid Feature Fusion. As stated in above, query features $F_Q \in \mathbb{R}^{H\times W\times C}$ are firstly down-sampled by spatial average pooling with strides $S = [s^1, ...s^n]$, and then followed by a 3×3 convolution layer that maintains the spatial dimensions unchanged. Assuming that the pooled query features are $\hat{F_Q} = \{f_Q^i\}_{i=1}^n$ with spatial dimensions $\{(\frac{H}{S^i}, \frac{W}{S^i})\}_{i=1}^n$ And the category-oriented features $F_C = \{f_C^i\}_{i=1}^n$ can be obtained by extending embedding vectors to corresponding dimensions. The correlation of support information and query images are established by

$$f_{fuse}^i = Conv^2(Cat(f_Q^i, f_C^i)), \quad i \in [1, n] \tag{3.5}$$

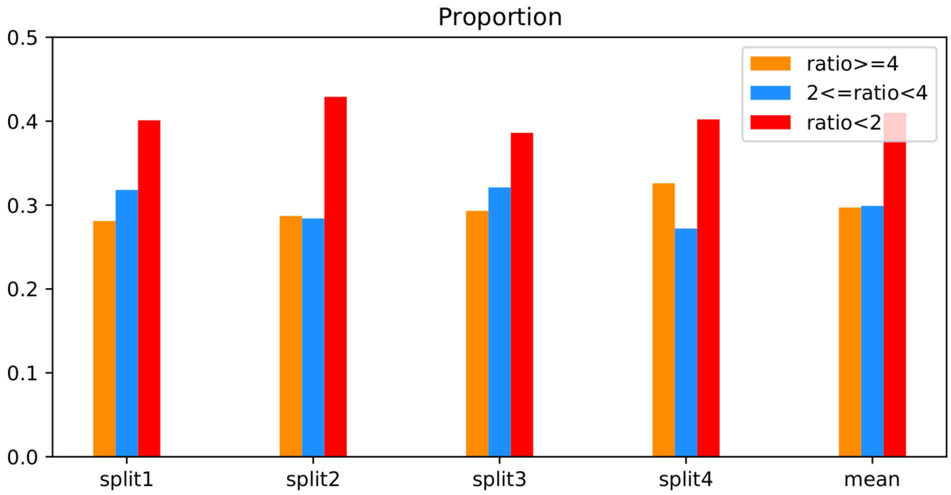


FIGURE 3.5. Histograms of scale variances between support masks and query masks.

where $Conv$ and Cat represents a 1×1 convolution and concatenation operation, respectively. f_{fuse}^{i} is the component of fused feature F_{fuse}.

Global Feature Extraction. Effectiveness of global contexts has been proved in many dense prediction tasks, *e.g.*, semantic segmentation (Peng et al., 2017; Yu et al., 2018) and salient object detection (Wang et al., 2018a; Zhang et al., 2018; Liu et al., 2019). To extract global context information, we introduce a global feature extraction module. Inspired by PSPNet (Zhao et al., 2017), the average pooling with different bin sizes are used to extract multiple pyramid features. As is shown in Figure 3.4(b), We use bin sizes of 1×1, 3×3 and 5×5. The pyramid features are further processed by a 1×1 convolution layer, and then directly up-sampled to original input sizes for global representation. Finally, we down-sample those features that contain global information to different scales for progressive feature integration, i.e., $F_{global} = \{f_{global}^{i}\}_{i=1}^{n}$.

Progressive Feature Integration. A direct way of integrating F_{fuse} and F_{global} of different scales is to match them at each level and then merge the features together. However, the variation of scales and appearances between support and query objects could result in the mismatch in a certain layers. To solve it, we propose a Progressive Feature Integration (PFI) to rectify the deviations. As is shown in Figure 3.4(c),It progressively constructs inner connections between adjacent feature maps of different levels. As a result, it yields more robust features for scale variations. More specifically, the proposed PFI incorporates recurrent connections from f_{fuse}^{i}, F_{global}^{i} as well as the output of previous stage $i-1$. The generated four level features can be represented as $F_{RFI} = \{f_{RFI}^{i}\}_{i=1}^{4}$. In stage i, the features f_{RFI}^{i} is obtained by:

$$f_{RFI}^{i} = Conv^{2}\ (Cat\ (f_i + Down(f_i)), f_{global}^{i}) \tag{3.6}$$

$$f_i = Cat(f_{fuse}^{i} + Pool(f_{RFI}^{i-1}) \tag{3.7}$$

Where $Pool$ denotes the average pooling operation with down-sampling rate 2×2. $Down$ is residual branches which enrich the identity by feeding it into average pooling layer followed by 3×3 convolution layer and then up-sampling it to the initial resolution. Note that the term $Pool(f_{RFI}^{i-1})$ is ingored when $i = 1$. Finally, a modified res-block is proposed to integrate features F_{RFI} of different scales by first concatenating and then passing them in a residual form.

3.3.3 Dynamic Predicting Module

The key of FSS is to fully exploit the information from support set annotations. Previous works (Tian, Shen, and Chen, 2020; Tian et al., 2022) generate the leaning parameters of K different mask heads for an image with K instances. Unlike traditional methods, our unique contribution lies in the introduction of the Dynamic Predicting Module (DPM). The primary purpose of DPM is to incorporate support information conditionally, enabling a more precise prediction of the segmentation mask. One of the distinguishing features of our DPM is its efficiency in parameter generation. Unlike instance segmentation methods (Tian, Shen, and Chen, 2020; Tian et al., 2022) that might continuously generate parameters, our DPM does so only once since only one mask needs to be predicted.

It's a well-established fact in the field that the deeper layers of neural networks tend to capture more high-level semantic features and category-specific information. Leveraging this, our model extracts category vectors specifically from the last layer of the support branch, ensuring a rich representation that captures intricate details relevant to our segmentation task.

Moreover, to facilitate the dynamic generation of learnable parameters for the mask prediction layer, we introduce a lightweight filter-generating network. This is a pivotal part of our methodology, enabling the adaptive and on-the-fly generation of parameters suited to the task at hand.

Lastly, the mask prediction layer is defined as a 1×1 convolution layer. This design choice simplifies the architecture, ensuring faster computation. Yet, it retains depth with its 2 channels, allowing the layer to handle the intricate task of mask prediction effectively. Supposing the input channels are 64, the mask predicting layer totally will contain 130 parameters ($weights = 64 \times 2$ and $bias = 2$). As is shown in Figure 3.2, we apply DPM on both multi-level features and the integrated features to predict intermediate segmentation maps and final segmentation maps.

To train our model, we introduce the deeply supervised learning and use the softmax cross-entropy loss for the main loss and auxiliary losses. Thus, the total loss $\mathcal{L}$ can be formally calculated by

$$\mathcal{L} = \mathcal{L}_{main} + \lambda \sum_{i=1}^{n} \mathcal{L}_{aux}^{i} \tag{3.8}$$

where $\mathcal{L}_{main}$ and $\mathcal{L}^{i}_{aux}$ are the main loss for final segmentation maps and auxiliary loss for intermediate segmentation maps respectively. λ is the loss weight for balance which is set to 1.

3.4 Experiments

3.4.1 Experimental settings

Datasets. Following previous works, public PASCAL-5^i (Shaban et al., 2017) and MS COCO-20^i (Nguyen and Todorovic, 2019) are used to verify the effectiveness of our method.

PASCAL-5^i is a widely-used dataset for FSS. It contains images from the PASCAL VOC 2012 (Everingham et al., 2010) with extra annotations from SDS (Hariharan et al., 2011). In PASCAL-5^i, the original 20 object categories are evenly divided into four folds for cross-validation. Specifically, three folds are used for training and the rest one is for testing. Following (Shaban et al., 2017), we use 1000 support-query pairs in each test.

MS COCO-20^i is modified from the more challenging dataset MS COCO (Lin et al., 2014). Similarly to the division in PASCAL-5^i, 80 categories are evenly divided into four splits. 60 categories are sampled for training and the remaining 20 are for testing in each split. Different from previous approaches (Nguyen and Todorovic, 2019; Yang et al., 2020a; Wang et al., 2020) that sampled 1000 support-query pairs in each split for testing, we sample 5000 pairs during testing to achieve more stable results.

Implementation Details. We exploit multi-level features from the ResNet-50 (He et al., 2016) pre-trained on ImageNet. To increase the receptive field, we also use the dilated version for ResNet-50 as previous works (Zhang et al., 2019b; Zhang et al., 2019a). The DRNet is optimized by SGD with the learning rate of 0.05 and momentum of 0.9. The learning rate decays with the "poly" policy (Chen et al., 2017a). We train our model for 100 epochs on PASCAL-5^i and 50 epochs on MS COCO-20^i with batch size 8.

Evaluation Metric. Mean-IoU and FB-IoU are employed as metrics for evaluating performances. Mean-IoU (Shaban et al., 2017) is the average of per-class foreground Intersection-over-Union (IoU) over all classes. For each category, the IoU is calculated by $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP and FN represent the number of true positives, false positives and false negatives. FB-IoU (Rakelly et al., 2018) is calculated by ignoring the object categories and averaging the IoU score of foreground and background over all test images.

3.4.2 Ablation Study

In this subsection, we first verify the effectiveness of the proposed key modules. Then we conduct more experiments on the configurations of PGM and DPM.

Effectiveness of key modules. To verify effects of each module in our method, we conduct ablation studies based on the following baseline. Table 3.1 shows the quantitative comparisons of 1-shot and 5-shot on PASCAL-5^i dataset.

The used baseline is based on CANet (Zhang et al., 2019b), which is a simple yet effective method for FSS. For fair comparison, we further simplify the architecture of CANet by removing the iterative optimization module and ASPP (Chen et al., 2017a). During testing, we also do not use multi-scale inputs. The result is shown in in Table 3.1 (1st row).

Based on the above baseline, we further verify the effectiveness of each module by adding different components. On the one hand, from the second row to the fourth row, it can be observed that the addition of each module increases results in terms of both Mean-IoU and FB-IoU. The most significant gain is brought by PGM and the averaged Mean-IoU improves from 54.57 to 58.77. One the other hand, by introducing NFFM, PGM and DPM, the performance can be further enhanced by a large margin. These results further demonstrate that our proposed modules could be complementary to each other.

TABLE 3.1. **Ablation studies** of the proposed architecture on PASCAL-5^i. Our baseline is the simplified version of CANet (Zhang et al., 2019b).

Baseline			Mean-IoU				
+NFFM	+PGM	+DPM	s-1	s-2	s-3	s-4	mean
			53.92	64.10	49.97	50.28	54.57
✓			54.81	64.45	50.19	51.17	55.16
	✓		60.79	68.01	52.51	53.88	58.77
		✓	54.57	64.73	51.04	50.62	55.24
✓	✓		61.13	67.83	53.27	54.40	59.16
✓	✓	✓	**61.99**	**68.87**	**53.74**	**55.02**	**59.91**

TABLE 3.2. **Comparison analysis of the PGM.**

Settings	Mean-IoU				
	s-1	s-2	s-3	s-4	mean
Baseline + PGM	60.79	**68.01**	**52.51**	**53.89**	**58.77**
w/o GFE	**60.84**	66.98	52.13	52.90	58.21
w/o PFI	59.47	66.23	51.84	52.34	57.57

Designing Choices of PGM To further understand the effect of PGM, we perform two additional experiments. Compared results are shown in Table 3.2. We alternatively remove the GFE and PFI. The results are shown in the 2nd row and 3rd row. We can see that the performance drops when compared with the complete module, demonstrating that the GFE and PFI are indispensable.

TABLE 3.3. **Comparisons of features selection for DPM.**

Layer	Mean-IoU				
	s-1	s-2	s-3	s-4	mean
Res-3	61.38	67.85	53.56	54.89	59.42
Res-4	**61.99**	**68.87**	**53.74**	**55.02**	**59.91**

TABLE 3.4. **Results with various input channels of mask predicting layer.** When Channels = 64, our model could achieve best results.

Channels	Mean-IoU				
	s-1	s-2	s-3	s-4	mean
32	60.72	68.41	**53.85**	54.67	59.41
64	**61.99**	68.87	53.74	55.02	**59.91**
128	61.34	**69.32**	53.32	54.46	59.61
256	61.21	67.91	53.55	**55.28**	59.49

Detailed Analysis of DPM. In this work, we exploit the DPM in two aspects: number of input channels in predicting heads and conditional layer to generate parameters of filters.

First, we verify the feature selection of DPM for generating parameters of the mask prediction layer. As shown in Table 3.3, choosing features of the Res-4 layer in ResNet-50 as input results in better performance. This demonstrates that it is more reasonable to choose deeper layers for object-level concepts. To clarify the influence of input channels at mask prediction layers, we change C_{input} in a range between 32 to 256. The results are shown in Table 3.4. As we can see, when $C_{input} = 64$, it delivers the best results. Thus, we keep it as a default number in other experiments.

We further prove the versatility of our DPM. Specifically, we re-implement PGNet (Zhang et al., 2019a) with DPM as a plug-in module. The comparison results with/without DPM are shown in Table 3.5. The DPM could increase the Mean-IoU of PGNet by 1.05%. From the results, we can observe that DPM is effective in proving performance with a considerable margin.

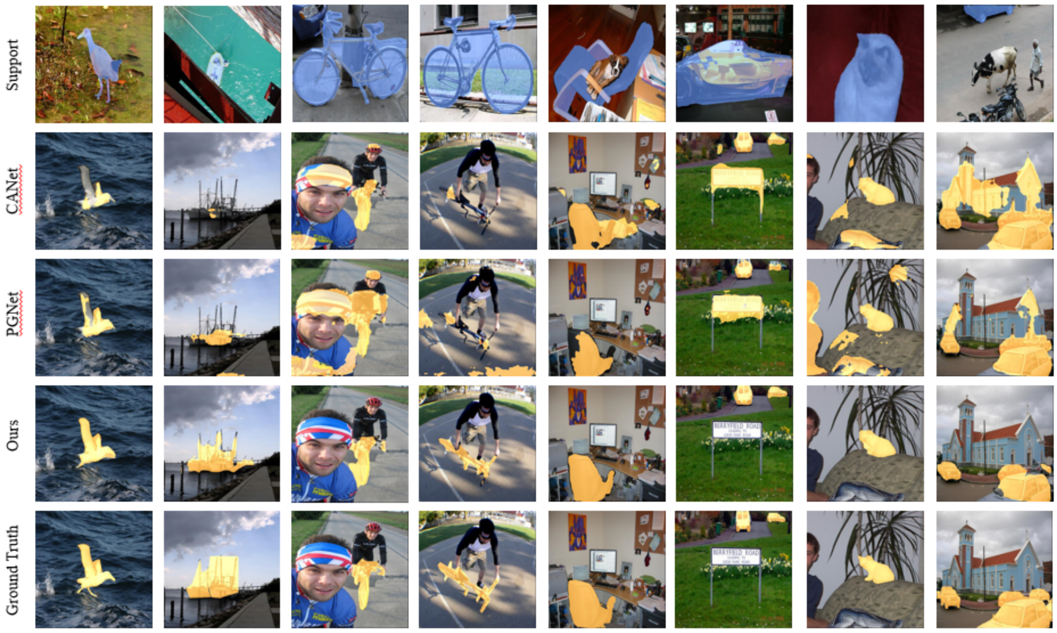
TABLE 3.5. **Results with the PGNet baseline** on PASCAL-5^i.

Method	Mean-IoU				
	s-1	s-2	s-3	s-4	mean
PGNet	55.07	63.92	47.15	50.93	54.27
PGNet+DPM	**55.61**	**64.79**	**47.31**	**51.64**	**54.84**

3.4.3 Comparison with State-of-the-arts

In this section, we compare the proposed method with other state-of-the-art methods including OSLSM (Shaban et al., 2017), AMP (Siam, Oreshkin, and Jagersand, 2019),

FIGURE 3.6. **Visual results** on PASCAL-5^i dataset. Even when suffering challenging scenes, *e.g.*, appearance and variations between query and support objects, our method can predict accurate segmentation maps



PANet (Wang et al., 2019a), FWB (Nguyen and Todorovic, 2019), RPMMs (Yang et al., 2020a), CANet (Zhang et al., 2019b), PGNet (Zhang et al., 2019a), PAP (Liu et al., 2020c) and DAN (Wang et al., 2020). For fair comparison, we use either the segmentation results or the codes provided by the corresponding authors.

Results on PASCAL-5^i. As shown in Table 3.6, our method significantly outperforms all previous methods under nearby all evaluation metrics. It convincingly demonstrates the effectiveness of the proposed method. Specifically, in the 1-shot setting, our method achieves 59.90% and 60.89% in terms of Mean-IoU and FB-IoU, respectively. It is worth mentioning that our method with the ResNet-50 backbone outperforms DAN (Wang et al., 2020) with the ResNet-101 backbone by 2.9% in the Mean-IoU and 1.4% in the FB-IoU. This result convincingly demonstrates the effectiveness of our method. Comparing the scores under 5-shot setting, our method also outperforms other algorithms by a large margin. Figure 3.6 provides some visual comparisons of our method with CANet (Zhang et al., 2019b) and PGNet (Zhang et al., 2019a). Comparing with the two previous works, we can observe that our method could generate accurate results under challenging scenes.

Results on MS COCO-20^i. We also quantitatively compare with previous methods on MS COCO-20^i dataset. The comparison results are shown in Table 3.7. Compared with the most recent works (Nguyen and Todorovic, 2019; Liu et al., 2020c; Yang et al., 2020a), our method could continuously achieve state-of-the-art results. Besides, there are some interesting observations. First, in both 1-shot setting and 5-shot setting, our methods outperform the very recent RPMMs (Yang et al., 2020a) by 10.8% and 7.8% in terms of average Mean-IoU scores. Second, our method is inferior to the RPMMs under split-1 subset. The reason may be that the split-1 subset is more complex in object parts. Under this case, RPMMs can perform better with prototype mixtures. However, our method delivers very comparable performances. Third, our method performs better than RPMMs on other four splits, which demonstrates the effectiveness of our method when dealing with various scenes. We also provide some visual examples. However, due to the limitation of space, we arrange them in supplemental materials.

3.5 Conclusion

In this paper, we propose a Deep Reasoning network (DRNet) to solve few-shot semantic segmentation. Different from previous methods, the learnable parameters of our proposed predicting layer are dynamically generated on support features. It effectively exploits the category-level information of deep layers in support branch. Besides, we develop a pooling-based guidance module to correlate multi-scale features of support and query branches. Our proposed non-local feature fusion module could help to fuse features of different levels in each branch. Extensive experiments on two public benchmarks demonstrate the effectiveness of our method.

TABLE 3.6. **Performance comparison of 1-shot and 5-way semantic segmentation on PASCAL-5^i dataset.** ‡ denotes that the model is evaluated with multi-scale inputs. Our method ranks first under most of these metrics.

Methods	Backbone	Mean-IoU(1shot)					FB-IoU	Mean-IoU(5shot)					FB-IoU
		s-0	s-1	s-2	s-3	mean	(1-shot)	s-0	s-1	s-2	s-3	mean	(5-shot)
OSLSM$_{\text{(BMVC'17)}}$(Shaban et al., 2017)	VGG-16	33.60	55.30	40.90	33.50	40.80	61.30	35.90	58.10	42.70	39.10	43.90	61.50
AMP-2$_{\text{(ICCV'19)}}$(Siam, Oreshkin, and Jagersand, 2019)	VGG-16	41.90	50.20	46.70	34.70	43.40	61.90	40.30	55.30	49.90	40.10	46.40	62.10
PANet$_{\text{(ICCV'19)}}$(Wang et al., 2019a)	VGG-16	42.30	58.00	51.10	41.20	48.10	66.50	51.80	64.60	59.80	46.50	55.70	70.70
FWB$_{\text{(ICCV'19)}}$(Nguyen and Todorovic, 2019)	VGG-16	47.04	59.64	52.51	48.27	51.90	-	50.87	62.86	56.48	50.09	55.08	-
RPMMs$_{\text{(ECCV'20)}}$(Yang et al., 2020a)	VGG-16	47.14	65.82	50.57	48.54	53.02	-	55.15	66.91	52.61	50.68	56.34	-
CANet$^{\ddagger}_{\text{(CVPR'19)}}$(Zhang et al., 2019b)	ResNet-50	52.50	65.90	51.30	51.90	55.40	66.20	55.50	67.80	51.90	53.20	57.10	69.60
PGNet$^{\ddagger}_{\text{(ICCV'19)}}$(Zhang et al., 2019a)	ResNet-50	56.00	66.90	50.60	50.40	56.00	-	57.70	68.70	52.90	54.60	58.50	-
PAP$_{\text{(ECCV'20)}}$(Liu et al., 2020c)	ResNet-50	47.83	58.75	53.80	45.63	51.50	-	58.39	67.83	**64.88**	56.73	61.69	-
FWB$_{\text{(ICCV'19)}}$(Nguyen and Todorovic, 2019)	ResNet-101	51.30	64.49	56.71	52.24	56.19	-	54.84	67.38	62.16	55.30	59.90	-
DAN$_{\text{(ECCV'20)}}$(Wang et al., 2020)	ResNet-101	54.70	68.60	**57.80**	51.60	58.20	71.90	57.90	69.00	60.10	54.90	60.50	72.30
Ours	ResNet-50	**61.99**	**68.87**	53.74	**55.02**	**59.91**	**72.52**	**64.15**	**69.92**	55.63	**57.35**	**61.76**	**74.13**

TABLE 3.7. **Performance comparison of 1-shot and 5-shot semantic segmentation on MS COCO-20^i dataset.**

Methods	Backbone	Mean-IoU(1shot)					FB-IoU	Mean-IoU(5shot)					FB-IoU
		s-0	s-1	s-2	s-3	mean	(1-shot)	s-0	s-1	s-2	s-3	mean	(5-shot)
FWB(ICCV'19)(Nguyen and Todorovic, 2019)	VGG-16	18.35	16.72	19.59	25.43	20.02	-	20.94	19.24	21.94	28.39	22.63	-
FWB(ICCV'19)(Nguyen and Todorovic, 2019)	ResNet-101	16.98	17.98	20.96	28.85	21.19	-	19.13	21.46	23.93	30.08	23.65	-
RPMMs(ECCV'20)(Yang et al., 2020a)	ResNet-50	29.53	**36.82**	28.94	27.02	30.58	-	33.82	**41.96**	32.99	33.33	35.52	-
Ours	ResNet-50	**33.64**	36.41	**33.92**	**31.15**	**33.78**	**57.92**	**37.29**	39.06	**38.54**	**36.22**	**37.78**	**60.20**

Statement of Authorship

Title of Paper	Few-shot Semantic Segmentation by Exploiting Dynamic and Regional Contexts	
Publication Status	□ Published □ Submitted for Publication	☑ Accepted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Accepted to ICME 2023	

Principal Author

Name of Principal Author (Candidate)	Yunzhi Zhuge		
Contribution to the Paper	Proposed the idea, made partial experiments and wrote the manuscript of the paper.		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	10/05/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Hongyu Gu		
Contribution to the Paper			
Signature		Date	12/05/2023

Name of Co-Author	Lu Zhang		
Contribution to the Paper			
Signature		Date	12/05/2023

Name of Co-Author	Jinqing Qi		
Contribution to the Paper			
Signature		Date	12/05/2023

Name of Co-Author	Huchuan Lu		
Contribution to the Paper	Discussion		
Signature		Date	12/05/2023

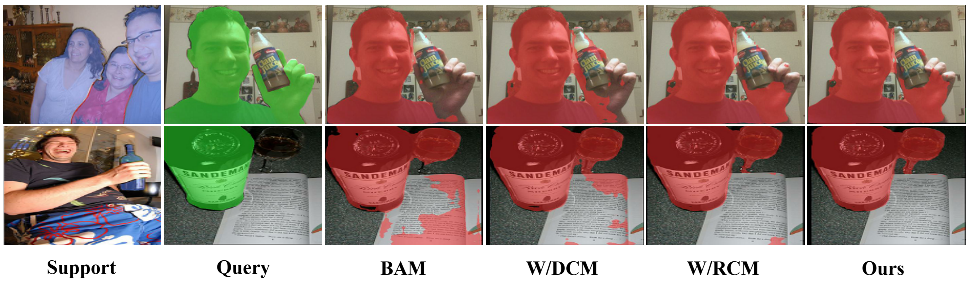


FIGURE 4.1. **Visual examples of ours** and BAM (Lang et al., 2022). From left to right: support images, query images, predictions of BAM, our baseline with DCM, our baseline with RCM and the final results.

Chapter 4

Few-shot Semantic Segmentation by Exploiting Dynamic and Regional Contexts

4.1 Introduction

Few-shot semantic segmentation is a challenging problem that aims to address the issue of scarce labeled data by learning to segment new object classes with only a few annotated samples. The existing prototype-based few-shot segmentation methods typically rely on the effective interaction between support and query images to learn a reliable segmentation model. However, modeling such interaction is a complex task, and the performance of these methods heavily depends on the quality of this interaction.

To tackle this issue, we propose a novel Dynamic and Regional Context Network (DRCNet) that achieves sufficient support-query interaction for accurate few-shot semantic segmentation. Our proposed approach leverages a Dynamic Context Module (DCM) to capture the spatial details in the query images. The DCM builds dynamic convolutions in local views, which complements the traditional global prototypes and forms multi-context interaction between support and query. This interaction leads to more accurate predictions on the query images.

In addition, we propose a Regional Context Module (RCM) to further improve the accuracy of our approach by modeling the prototypes for ambiguous regions and excluding the background and ambiguous objects in query images. Our experimental results on Pascal-5i and COCO-20i datasets demonstrate that our proposed DRCNet significantly outperforms state-of-the-art methods. Our approach provides a promising direction for accurate few-shot semantic segmentation, and our proposed modules can be readily integrated into existing models to enhance their performance.

4.2 Background

In recent years, deep learning has brought significant improvements to the field of semantic segmentation. However, one of the major challenges in training segmentation models lies in the dependence on large amounts of human annotated data. As is known to us all, annotating a dataset for semantic segmentation is a time-consuming and resource-intensive process, and it becomes even more challenging when the dataset contains a large number of object categories or a high level of intra-class variability.

Semi-supervised semantic segmentation has been proposed as a solution to this problem, which is based on the assumption that a large amount of unlabeled data is available in addition to a small number of labeled data. These approaches aim to learn from both labeled and unlabeled data to improve the segmentation performance. However, the requirement of labeled data is still a bottleneck in these approaches, and they are prone to generalizing poorly on novel categories.

To address this limitation, few-shot semantic segmentation has emerged as a promising solution, the purpose of which is to learn a model that can quickly adapt to new object categories with only a few annotated examples as the prompt. In other words, it allows the model to generalize to novel categories without the necessity of a large amount of labeled data. The goal is to leverage prior knowledge learned from the base categories to efficiently segment novel categories.

While current methods have produced impressive results in few-shot segmentation, they are not without limitations. One such limitation is the use of Masked Average Pooling (MAP) operation in the prototypical learning framework, which can lead to the destruction of spatial structure in the feature space. This, in turn, can have a negative impact on mask quality, particularly for objects with large shape variations. Another challenge that must be addressed is the issue of background noise and ambiguous regions present in query images. To tackle this issue, it is essential to develop a technique that can effectively model regional context information while simultaneously excluding background and ambiguous objects.

To alleviate the aforementioned problems, we propose a novel approach known as the Dynamic and Regional Context Network (DRCNet) for precise Few-shot Semantic Segmentation (FSS). DRCNet is designed to produce robust representations for both the target objects and ambiguous regions. The approach is based on the prototypical learning framework, which consists of two main components: the Dynamic Context Module (DCM) and the Regional Context Module (RCM). To begin with, DCM is employed to capture intricate spatial details by utilizing dynamic convolution to interact with both the support and query features. Dynamic kernels are learned from the support features to extract spatial information from the query features. The resultant features, combined with the query feature and traditional global prototype via MAP, are integrated to provide multi-context support-query interaction. The resulting features are then passed to a decoder to generate an initial prediction. In addition to DCM, we introduce the RCM, which is responsible for dealing with ambiguous regions in query images. This module calculates an uncertainty map to create prototypes that can effectively eliminate noise interference in query features and refine initial masks, resulting in more precise and accurate predictions. By leveraging the strengths of both DCM and RCM, DRCNet can overcome the challenges of few-shot semantic segmentation and produce more reliable results.

To summarize, our contributions are as follows:

- We propose a novel method called Dynamic and Regional Context Network (DRCNet) for accurate Few-shot Semantic Segmentation (FSS) by learning robust prototypes for both target objects and ambiguous regions.
- We introduce a Regional Context Module (RCM) to effectively capture and eliminate complex background and interference objects that belong to other categories.
- We propose a Dynamic Context Module (DCM) to model local interaction between query and support samples, enabling DRCNet to capture fine spatial details.
- Extensive experiments on PASCAL-5^i and COCO-20^i demonstrate the superior performance of the proposed DRCNet compared to state-of-the-art methods.

4.3 Our approach

4.3.1 Problem Setting

Few-shot Semantic Segmentation (FSS) is a challenging task that aims to accurately segment objects from novel classes using only a few annotated samples. To accomplish this, FSS uses a training set D_{tr} and a testing set D_{ts}, where the base classes and the novel classes are disjoint, i.e., $C_{tr} \cap C_{ts} = \emptyset$. Both D_{tr} and D_{ts} are composed of multiple episodes or sub-tasks, where each episode includes a support set S and a query set Q from the same class.

In the k-shot FSS scenario, the support set S can be represented as $S = \{(x_i^s, m_i^s)\}_{i=1}^{K}$, where xi^s and m_i^s) represent the support image and mask, respectively, and k is the

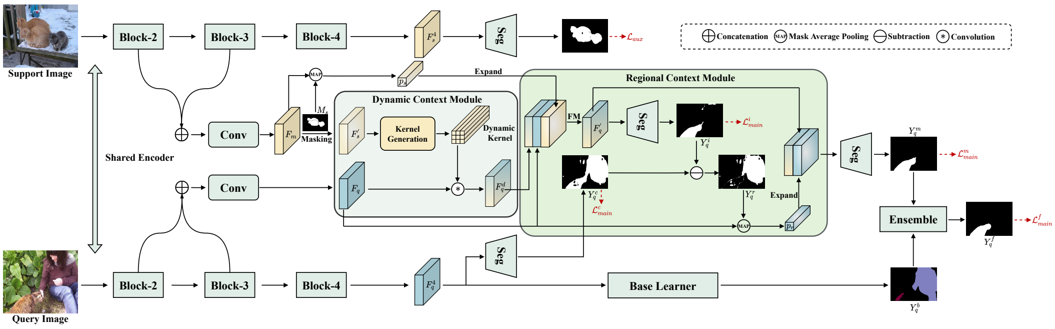


FIGURE 4.2. **Overall architecture of the proposed DRCNet.** Given the support and query features, the Dynamic Context Module (DCM) is used to perform their local interaction via dynamic convolution. The generated features together with the traditional MAP based prototypes and query feature are concatenated to generate an initial mask via decoder. Then, a Regional Context Module (RCM) is proposed to mine the prototype of ambiguous regions to further distill the noises of initial mask.

number of samples. Similarly, the corresponding query set can be formulated as $Q = \{x^q, m^q\}$. During training, an episode (S, G) is sampled, and the model takes S and x^q as input to produce a binary mask $\hat{m}^q$. The model is supervised with binary cross-entropy (BCE) loss to update its weights. Specifically, given that the resolution of predicted binary mask $\hat{m}^q$ and ground-truth m^q is $H \times W$, the BCE loss can be calculated as:

$$BCE(\hat{m}^q, m^q) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (-m_{ij}^q \cdot \log(\hat{m}_{ij}^q - (1 - m_{ij}) \cdot \log(1 - \hat{m}_{ij}^q - (1 - m_{ij}))) \tag{4.1}$$

Once the training process is done, the meta-testing is performed by taking randomly sampled episodes from D_{ts} for evaluation.

4.3.2 Overview

We introduce a novel approach, the Dynamic and Regional Context Network (DRCNet), for few-shot semantic segmentation, which aims to segment objects from novel classes using only a few annotated samples. The architecture of our proposed network is shown in Figure4.2. Here, we elaborate our work on the 1-shot setting for simplicity. The DRCNet is built upon the prototype-based framework and consists of two key sub-modules, namely Dynamic Context Module (DCM) and Regional Context Module (RCM). We employ ResNet-50 (He et al., 2016) and VGG-16 (Simonyan and Zisserman, 2014) as our backbone to extract multi-level features for both support and query images, which are denoted as $\{F_s^l\}_{l=1}^L$ and $\{F_q^l\}_{l=1}^L$, respectively.

To accomplish the interaction between support feature and query feature, we introduce the DCM, which uses dynamic convolution to capture the spatial details. Specifically, the kernels are produced from the support features, and applied to the query features

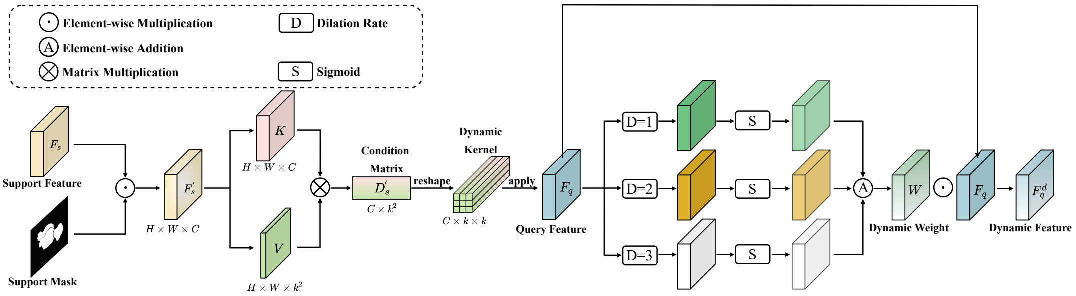


FIGURE 4.3. **The architecture of the dynamic context module.** The DCM takes the input as support feature, mask and query feature, and aims to perform sufficient interaction between them. First, the target objects are emphasized in support feature by support mask, which are decoupled into key and value to form dynamic kernel. Then, the dynamic kernels are applied to query features using dynamic convolution to model the pixel-wise interaction between support region and query image. Finally, a multi-context attention is applied to the query feature to produce the output dynamic feature.

to extract more informative features. These generated features, combined with the original global prototypes and query features, are fed to a decoder to generate an initial mask. The dynamic convolution allows our model to adapt to the target object in the query image, resulting in more accurate segmentation performance.

Moreover, to eliminate the interference from the background, we propose the RCM to model the prototype of those ambiguous regions. The RCM is designed to learn a more robust representation of the ambiguous regions and exclude them from the segmentation process. This is achieved by modeling the prototype of the ambiguous regions based on a calculated uncertainty map. The RCM suppresses the interference in the query features and refines the initial masks, thus could obtaining more accurate and robust predictions.

We explain the details of DCM and RCM in the following sections, and demonstrate the effectiveness of our proposed approach through comprehensive experiments.

4.3.3 Dynamic Context Module

In few-shot semantic segmentation, prototype-based methods (Zhang et al., 2019b; Tian et al., 2020; Mao, 2022) have been commonly used to compress feature vectors that represent target objects. The Masked Average Pooling (MAP) on support features enables the prototypes to capture the global context of target regions, which helps in accurately recognizing truly-matching objects in query images. However, the accuracy of these models, particularly on objects with complex shape variations, tends to decline rapidly due to the lack of adequate spatial details that come with average pooling. To tackle this problem, we propose a Dynamic Context Module (DCM) that models the local context between support and query features. The DCM complements the global prototypes by capturing fine details in query images. The architecture of DCM is illustrated in Figure 4.3 to provide a clearer understanding of the mechanism.

The DCM uses dynamic convolution which enables the model to capture spatial details in the support features. The kernels generated by the DCM from the support features facilitate the interaction between support feature and query feature. This interaction allows for the capture of fine-grained details in the query images, which are crucial for accurate segmentation results. The generated features are then combined with the original global prototypes and query features and fed to a decoder to generate an initial mask. This process allows the model to eliminate background noise and focus on the target objects. In the following sections, we will elaborate on the details of the DCM and the Regional Context Module (RCM) that complement the DCM to improve the performance in few-shot semantic segmentation.

Following previous works (Zhang et al., 2019b; Tian et al., 2020), we use the mid-level features to form the support and query features respectively as:

$$F_s = Conv_{1\times 1}(Cat(F_s^2, F_s^3)) \tag{4.2}$$

$$F_q = Conv_{1\times 1}(Cat(F_q^2, F_q^3)) \tag{4.3}$$

where $Conv_{1\times 1}$ indicates convolution with kernel size as 1×1. F_s^2 and F_s^3 are extracted features of support image on the second and third stage of backbone, and so as F_q^2 and F_q^3. Cat denotes the concatenation along channel dimensions. F_s and F_q are support and query features, which will be fed to DCM. To emphasize the foreground regions, the support feature F_s is first multiplied with the corresponding mask M_s by:

$$F_s^{'} = F_s \odot M_s \tag{4.4}$$

where $\odot$ is the Hadamard product. Note that the support mask M_s is resized to the same resolution as F_s. This allows for better alignment between the features and the mask, which is important for accurate segmentation. Instead of using a fully connected layer to generate dynamic kernels from the support feature $F_s^{'}$, we utilize a more efficient method based on matrix multiplication (Liu et al., 2020a). Specifically, we first apply two independent 1×1 convolutions to transform $F_s^{'}$ into the key $K \in \mathbb{R}^{H\times W\times C}$ and value $V \in \mathbb{R}^{H\times W\times k^2}$. This approach efficiently explores the categorical information in support features from both aspects by modulating $F_s^{'}$ from the perspective of channel using the key feature K, and capturing the global spatial distributions of $F_s^{'}$ with the value feature V. This allows the dynamic kernels to be generated in a more effective and computationally efficient manner, leading to improved performance in segmentation tasks. To generate the dynamic kernels, we first reshape the key feature and value feature to $K^{'} \in \mathbb{R}^{N\times C}$ and $V^{'} \in \mathbb{R}^{N\times k^2}$, where $N = H \times W$. Then, the condition matrix $D_s^{'} \in \mathbb{R}^{C\times k^2}$ is obtained by:

$$D_s^{'} = (K^{'})^T V^{'} \tag{4.5}$$

Finally, we reshape the condition matrix $D_s^{'} \in \mathbb{R}^{C \times k^2}$ to $D_s \in \mathbb{R}^{C \times k \times k}$ to obtain the dynamic kernels.

After applying the generated kernels D_s to the query features, we still need to address the scale inconsistency between the support and query objects. To tackle this issue, we establish multi-scale contextual relations by creating a triplet of dynamic kernels with various dilation rates $R \in 1, 2, 3$. This allows us to capture objects at different scales, which is crucial for accurate recognition. However, rather than directly obtaining the output by applying these kernels to the query features, we first generate weighting maps W_1, W_2, and W_3, each of which corresponds to a different dilation rate. These maps are then averaged to produce the final weighting map W. This approach allows us to combine information from multiple scales, resulting in more robust and accurate feature representation. Finally, we enhance the feature representation even further through a feature matching process (Zhang et al., 2019b), which leads to the generation of $\tilde{F}_q$ by:

$$F_q' = Conv_{1\times1}(Cat(F_q, C_q, F_q^D)) \tag{4.6}$$

where C_q is traditional prototype by MAP. Finally, We use F_q' to predict an initial mask $Y_q^i \in \mathbb{R}^{H \times W \times 1}$ via a segmentation head:

$$Y_q^i = Conv_{1\times1}(Conv_{3\times3}(ASPP(F_q^{'}))) \tag{4.7}$$

where $ASPP$, $Conv_{3\times3}$ and $Conv_{1\times1}$ are the atrous spatial pyramid pooling layer, 3×3 convolution layer and 1×1 convolution layer, respectively. This refined feature representation is a critical step in the matching process and helps to improve the accuracy of the final segmentation result.

4.3.4 Regional Context Module

The proposed feature F_q' generated using DCM has been effective in building pixel-wise relationships between query images and target objects from support images. However, in complex scenes with interference objects or backgrounds from other classes, relying solely on the support target regions may not always provide sufficient discrimination. To address this issue, we introduce RCM, which is designed to mine ambiguous regions in query images.

In a previous study (Liu et al., 2022c), the authors proposed using proxy masks to represent the support foreground, background, and distracting regions. While this approach has been effective in some cases, it has several fatal drawbacks. One such drawback is the inconsistency between the support and query images, which can lead to misleading information and result in an enlarged discrepancy of distracting. For example, in a support image scene, a cat may be lying on a sofa, while in the corresponding query image, another cat may be held by a person. This inconsistency may prevent the model from extracting common cues between the support and query images, and such misleading information may accumulate as the discrepancy of distracting being enlarged. Unlike DCPNet, which relies on support images, our method focuses on the query image itself, avoiding inconsistencies between the support and query images. By doing so, RCM can extract the most relevant information from the query image, thus improving the overall performance of the model. Therefore, we propose RCM to mine the ambiguous regions in query images to complement the target regions in support images for better discrimination.

Given the high-level query feature F_q^4, we directly attach a segmentation head behind to obtain a coarse mask $Y_q^c \in \mathbb{R}^{H \times W \times 1}$.

$$Y_q^c = Conv_{1\times1}(Conv_{3\times3}(F_q^4)) \tag{4.8}$$

As shown in Figure 4.2, the coarse mask is inaccurate with complex noise from background.

After obtaining the initial and coarse predictions Y_q^i and Y_q^c, respectively, we refine the query feature $\tilde{F}q$ by leveraging the differences between the two predictions. Specifically, while Yq^i is designed to capture the target objects, Y_q^c provides a more global view of the scene. Therefore, we can exploit the contrast between these two predictions to extract more discriminative features that can help to distinguish foreground objects from complex backgrounds.

In other words, the differentiation between Y_q^i and Y_q^c contains external regional contexts that the model can learn from to mine more informative features. By doing so, the refined query feature can capture the subtle details of the foreground objects that are difficult to extract from the initial prediction. Additionally, the model can learn to differentiate between foreground and background more effectively, as the external contexts provide a more comprehensive view of the scene. This refinement process thus improves the overall performance of the model by enhancing the discriminative power of the query feature.

To be specific, we first obtain the regional compensate mask $Y_q^r \in \mathbb{R}^{H \times W \times 1}$ by:

$$Y_q^r = \left| Y_q^c - Y_q^i \right| \tag{4.9}$$

The following step involves masked average pooling with Y_q^r on the query feature F_q. This pooling operation allows us to obtain a regional compensate prototype vector $p_r \in \mathbb{R}^{1 \times 1 \times C}$, which we then expand to $p_r^{'} \in \mathbb{R}^{H \times W \times C}$. The expanded vector $p_r^{'}$ contains information about the regional context of the query image, which can help us to better distinguish foreground objects from complex backgrounds.

With the expanded vector $p_r^{'}$ in hand, we combine it with the activated query feature $F_q^{'}$. This combination is then passed through a segmentation head, allowing us to obtain the region refined mask $Y_q^m \in \mathbb{R}^{H \times W \times 1}$:

$$Y_q^m = Conv_{1\times1}(Conv_{3\times3}(ASPP(Cat(F_q^{'}, p_r^{'})))) \tag{4.10}$$

Note that the segmentation head in RCM is not shared with that in DCM. This mask provides us with information about the foreground objects in the query image, which we can use to better localize and classify these objects. By using the regional compensate prototype vector in combination with the query feature, we are able to improve the overall accuracy of our model in identifying foreground objects in complex scenes. Finally, we use the ensemble module (Lang et al., 2022) to obtain our final predict mask Y_q^f by:

$$Y_q^f = F_{ensemble}(Y_q^m, Y_b) \tag{4.11}$$

where Y^b is the prediction of base learner and $F_{ensemble}$ is a lightweight mask adjust module to effectively integrate those two signals. Specifically, $F_{ensemble}$ comprises two 1×1 convolutional operations. One operation refines the coarse results from the meta learner, and the other operation fuses the outputs of the two learners using designated initial parameters.

4.3.5 k-shot Setting

In the case of k-shot learning where there are more annotated support images available, previous methods have mainly relied on either averaged support prototypes to guide query features or direct forward of k times. However, (Lang et al., 2022) has demonstrated that both these methods are likely to produce sub-optimal results. To overcome this issue, we adopt the adaptive weighting approach proposed by BAM to assign importance to different support images. Specifically, we concatenate the adjustment factors of support samples to form a unified vector, which is then processed through two fully connected layers to generate fusion weights. By doing this, we can effectively capture the importance of each support image and use this information to guide the learning process.

Once we obtain the fusion weights, the ensemble can be achieved by performing a weighted summation between feature maps and their corresponding fusion weights. This approach enables taking into account the contribution of each support image in a more effective way, leading to better performance in k-shot learning scenarios. By assigning different weights to different support images, we can capture the nuances and complexities of the dataset more effectively, which can be particularly beneficial in scenarios where there is significant variability between different support images. Overall, our approach provides a more flexible and robust way to handle k-shot learning tasks, and we believe it will prove to be a valuable addition to the existing literature in this area.

4.3.6 Training Loss

Our training process involves using both binary cross-entropy (BCE) loss and cross-entropy (CE) loss. To ensure an optimal learning outcome, the training is divided into two distinct stages: the base-training stage and the meta-training stage. During the base-training stage, a standard supervised learning paradigm is employed to train the base learner using CE loss. This is done by comparing the base prediction Y^b to the ground-truth M^b for training the base learner as:

$$\mathcal{L}_{base} = CE(Y^b, M^b) \tag{4.12}$$

Moving on to the meta-training stage, we utilize an episodic learning paradigm to train both the joint meta learner and ensemble module. In this stage, we use a combination of BCE losses to enhance the learning process, which encourage the model to focus on the foreground objects and avoid confusing them with the background or interfering objects. Specifically, we apply four BCE losses to supervise the prediction of coarse prediction Y_q^c, initial prediction Y_q^i, meta prediction Y_q^m and the final prediction Y_q^f, which can be formulated as:

$$\begin{aligned} \mathcal{L}_{main} = &BCE(Y_q^c, M_q) + BCE(Y_q^i, M_q) \\ &+ BCE(Y_q^m, M_q) + BCE(Y_q^f, M_q) \end{aligned} \tag{4.13}$$

where M_q is the ground truth of query objects. By doing this, the ensemble module learns to make a more accurate prediction of the foreground objects.

We also introduce additional constraints on the predictions of support masks Y_s by:

$$\mathcal{L}_{aux} = \frac{1}{K}\sum_{k=1}^{K} BCE(Y_s^{ck}, M_s^k) \tag{4.14}$$

where k is the amount of available support pairs. Thus, the total training loss for each episode is:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda \mathcal{L}_{aux} \tag{4.15}$$

where λ is a trade-off and set to 1.0 empirically.

4.4 Experiments

4.4.1 Datasets and Evaluation Metrics.

Datasets. The evaluation and comparison of our method with other state-of-the-art techniques are carried out on two benchmarks: Pascal-5^i (Shaban et al., 2017) and

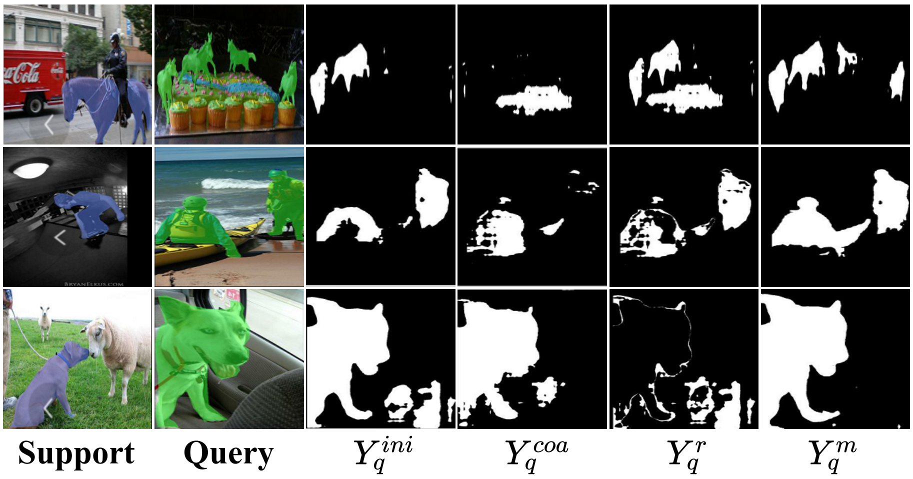


FIGURE 4.4. **Ablation study on regional context module under 1-shot setting.** From left to right: support images, query images, initial predictions, coarse predictions, regional compensate masks and regional refined predictions.

COCO-20^i (Kang et al., 2019). Pascal-5^i is derived from PASCAL VOC 2012 (Everingham et al., 2010) along with its extension dataset SDB (Hariharan et al., 2011). It comprises a total of 20 classes, which are distributed equally among four folds, indicating that each fold consists of five categories. On the other hand, COCO-20^i is a larger benchmark, modified from the MSCOCO (Lin et al., 2014) dataset, with 80 classes. Similarly to Pascal-5^i, the categories in COCO-20^i are partitioned into four folds, and each fold contains 20 classes. For evaluating the performance of our model on both Pascal-5^i and COCO-20^i datasets, we have used cross-validation technique.

Evaluation metrics. We conduct all the experiments using the Pytorch Toolkit. Following BAM (Lang et al., 2022), the training process is divided into two stages, i.e., pre-training of base learner using standard supervised learning paradigm and meta-training of joint meta learner and ensemble module using episodic learning paradigm. For base learner, we continue to use PSPNet (Zhao et al., 2017) to predict segmentation results on base classes. As for the meta-training, we freeze the parameters of encoder and base learner and optimize the rest of the network with SGD for 100 epochs and 50 epochs on Pascal-5^i and COCO-20^i respectively with initial learning rate of 5e-3 and 2.5e-3. All the experiments are run on NVIDIA RTX TITAN GPUs.

4.4.2 Ablation Studies

Ablation study on effectiveness of different components. In order to analyze the effectiveness of each component in DRCNet, we conducted a series of experiments on Pascal-5^i using Resnet-50 as the feature extractor. The results of these experiments are shown in Figure 6.2. It is clear from the results that both components contribute to improving the overall performance of the network. Specifically, we observed that (i) the ensemble operation led to a significant increase in mIoU, with the score increasing

Table 4.1. **Ablation study on effectiveness of different components.**

Ensemble	DCM	RCM	split0	split1	split2	split3	mean
			65.48	71.34	65.38	58.82	65.25
✓			68.49	73.20	66.40	60.89	67.24
✓	✓		68.83	73.89	66.94	61.09	67.69
✓		✓	70.02	74.51	67.80	61.72	68.50
✓	✓	✓	**70.36**	**74.68**	**67.87**	**61.98**	**68.72**

Table 4.2. **Ablation study on k-shot fusion.**

Method	mIoU	FB-IoU
1-shot baseline	68.63	80.10
Mask vote	69.21	80.41
Mask average	69.24	80.89
Feature average	71.49	82.55
Reweighting	**71.96**	**82.93**

by 3.05%; (ii) when DCM and RCM were included, the mIoU increased by 3.74% and 5.00%, respectively. Furthermore, when both DCM and RCM were combined, we were able to achieve an even greater improvement of 5.32% in mIoU. These results demonstrate that the combination of these different components is crucial for achieving optimal performance in DRCNet.

Ablation study on k-shot fusion. DRCNet adopts reweighting, an early fusion strategy for k-shot setting which learns an adjusting value ψ to allocate weights for different support samples. We compare our reweighting with two late fusion strategies,i.e.,mask voting (Min, Kang, and Cho, 2021) and mask average (Zhang et al., 2019b), and feature average (Tian et al., 2020) which is another common early fusion way. Our reweighting performs best among all the compared methods.

Ablation study on regional context module. To provide a clearer understanding of the impact of the regional context module(RCM), we performed a comparison of the predictions obtained at different stages within RCM. As shown in Figure 4.4, our observations suggest that RCM has a crucial role in addressing missing areas(1st row), enhancing the completeness of predictions(2nd row), and eliminating ambiguity in the predicted results(3rd row). In other words, the RCM component of our model is essential for improving the overall performance and accuracy of the segmentation task.

4.4.3 Comparison with State-of-the-Art Methods

Pascal-5^i. Table 4.4 presents the performance comparisons of our DRCNet with other state-of-the-art methods on Pascal-5^i. Our method obtains the best results in terms of both mean intersection over union (mIoU) and foreground-background IoU (FB-IoU) in various situations, i.e., using various backbones (VGG-16 and Resnet-50) and under

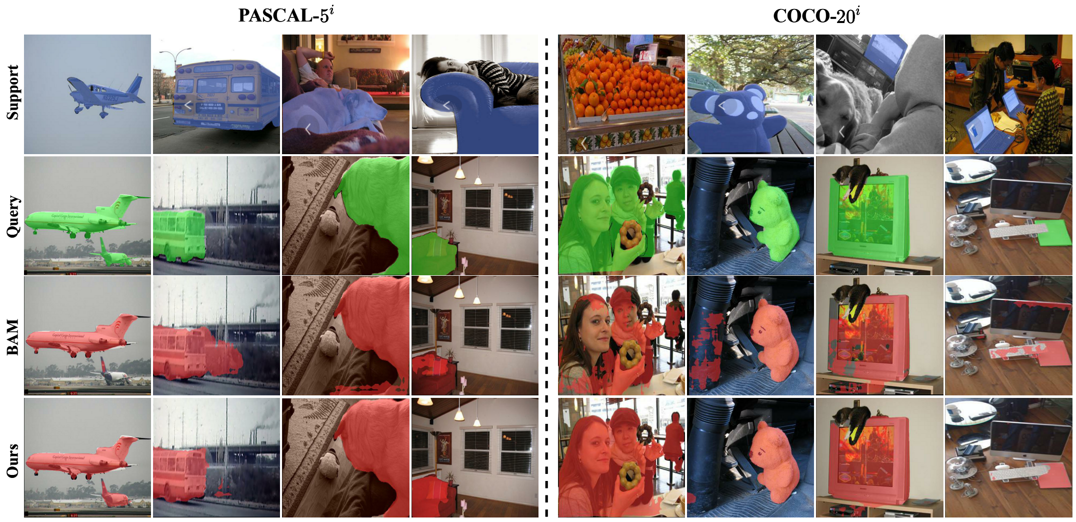


FIGURE 4.5. **Qualitative comparisons of our method with BAM under 1-shot setting.** We show the results of Pascal-5^i on left part and COCO-20^i on the right.

both 1-shot and 5-shot settings. Specifically, under the 1-shot few-shot segmentation (FSS) setting, the mIoUs of our method are consistently better than other methods on all four folds. Moreover, our method achieves the mean mIoU of 65.20 and 68.63 using VGG-16 and Resnet-50, respectively, outperforming previous state-of-the-art methods by 1.46% and 1.21%. When it comes to the 5-shot FSS setting, our method still outperforms the previous best method by 1.46%.

In addition, Table 4.3 compares the average FB-IoU scores on PASCAL-5^i. Our method also outperforms other methods by a remarkable margin, which indicates that our method can better capture the foreground and background regions of the objects. Overall, the results demonstrate the effectiveness of our DRCNet in addressing the few-shot segmentation problem on Pascal-5^i.

COCO-20^i. In Table 4.5, we compare with state-of-the-art methods on averaged mean-IoUs using Resnet-50 as the backbone. It is evident from our results that our proposed method outperforms existing approaches significantly, particularly in the 1-shot setting where we achieve an improvement of 4.59% over the closest competitor.

Qualitative comparison. To qualitatively compare our method with BAM (Lang et al., 2022), we present predictions on Pascal-5^i and COCO-20^i datasets under the 1-shot setting. The predictions demonstrate that our DCRNet is capable of capturing multiple contexts through DCM and RCM, thereby reducing the impact of complex backgrounds and distracting objects (2nd, 5th, and 8th columns). Moreover, our method outperforms BAM (Lang et al., 2022) in terms of predicting structurally consistent results (4th and 7th columns). Thus, based on our observations, we can conclude that DCRNet has superior performance in comparison to BAM (Lang et al., 2022) under the 1-shot setting.

Table 4.3. **Averaged FB-IoU results of 4 folds on PASCAL-5^i.**

Backbone	Methods	FB-IoU (%)	
		1-shot	5-shot
VGG16	PFENet	72.00	72.30
	HSNet	73.40	76.60
	DPCN	73.70	77.20
	NTRE	73.10	74.20
	BAM	77.26	81.10
	DRCNet	**78.11**	**81.79**
ResNet50	PFENet	73.30	73.90
	HSNet	76.70	80.60
	DPCN	78.00	80.70
	NTRE	77.00	78.40
	BAM	79.71	82.18
	DRCNet	**80.10**	**82.93**

TABLE 4.4. **Class mIoU results of four folds on PASCAL-5^i.**The 'Mean' column denotes the averaged class mIoU scores of all the four folds. **Bold** and underline indicate the best and second best results.

Backbone	Methods	1-Shot					5-Shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
VGG-16	RPMM (Yang et al., 2020a)	47.14	65.82	50.57	48.54	53.02	50.00	66.46	51.94	47.64	54.01
	PFENet(Tian et al., 2020)	56.90	68.20	54.40	52.40	58.00	59.00	69.10	54.80	52.90	59.00
	MMNet(Wu et al., 2021)	57.10	67.20	56.60	52.30	58.30	56.60	66.70	63.60	56.50	58.30
	HSNet (Min, Kang, and Cho, 2021)	59.60	65.70	59.60	54.00	59.70	64.90	69.00	64.10	58.60	64.10
	DPCN (Liu et al., 2022a)	58.90	69.10	63.20	55.70	61.70	63.40	70.70	68.10	59.00	65.30
	NTRE (Liu et al., 2022c)	57.70	67.60	57.10	53.70	59.00	60.30	68.00	55.20	57.10	60.20
	BAM (Liu et al., 2022c)	63.18	70.77	66.14	57.53	64.41	67.36	73.05	**70.61**	64.00	68.76
	Ours	**64.49**	**71.87**	**66.50**	**58.52**	**65.35**	**69.22**	**74.42**	70.40	**64.15**	**69.55**
ResNet-50	RPMM (Yang et al., 2020a)	55.15	66.91	52.61	50.68	56.34	56.28	67.34	54.52	51.00	57.30
	PFENet(Tian et al., 2020)	61.70	69.50	55.40	56.30	60.80	63.10	70.70	55.80	57.90	61.90
	SCL(Zhang, Xiao, and Qin, 2021)	63.00	70.00	56.50	57.70	61.80	64.50	70.90	57.30	58.70	62.90
	ASGNet(Li et al., 2021)	58.84	67.86	56.79	53.66	59.29	63.66	70.55	64.17	57.38	63.94
	SAGNN(Xie et al., 2021b)	64.70	69.60	57.00	57.20	62.10	64.90	70.00	57.00	59.30	62.80
	MLC(Yang et al., 2021a)	59.20	71.20	65.60	52.50	62.10	63.50	71.60	71.20	58.10	66.10
	HSNet (Min, Kang, and Cho, 2021)	64.30	70.70	60.30	60.50	64.00	70.30	73.20	67.40	67.10	69.50
	DPCN (Liu et al., 2022a)	65.70	71.60	69.10	60.60	66.70	70.00	73.20	70.90	65.50	69.90
	NTRE	65.40	72.30	59.40	59.80	64.20	66.20	72.80	61.70	62.20	65.70
	BAM (Lang et al., 2022)	68.97	73.59	67.55	61.13	67.81	70.59	75.05	**70.79**	67.20	70.91
	Ours	**70.36**	**74.68**	**67.87**	**61.98**	**68.72**	**72.36**	**76.54**	70.66	**68.26**	**71.96**

TABLE 4.5. **Class mIoU results of four folds on COCO-20^i.** The 'Mean' column denotes the averaged class mIoU scores of all the four folds. **Bold** and underline indicate the best and second best results.

Backbone	Methods	1-Shot					5-Shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
ResNet-50	RPMM (Yang et al., 2020a)	29.50	36.80	28.90	27.00	30.60	33.80	42.00	33.00	33.30	35.50
	RePRI (Boudiaf et al., 2021)	31.20	38.10	33.30	33.00	34.00	38.50	46.20	40.00	43.60	42.10
	CWT (Lu et al., 2021)	32.20	36.00	31.60	31.60	32.90	40.10	43.80	39.00	42.40	41.30
	CMN (Xie et al., 2021c)	37.90	44.80	38.70	35.60	39.30	42.00	50.50	41.00	38.90	43.10
	MMNet (Wu et al., 2021)	34.90	41.00	37.20	37.00	37.50	37.00	40.30	39.30	36.00	38.20
	HSNet (Min, Kang, and Cho, 2021)	36.30	43.10	38.70	38.70	39.20	43.30	51.30	48.20	45.00	46.90
	NTRE (Liu et al., 2022c)	36.80	42.60	39.90	37.90	39.30	38.20	44.10	40.40	38.40	40.30
	BAM (Lang et al., 2022)	43.41	50.59	47.49	43.42	46.23	49.26	54.20	51.63	49.55	51.16
	Ours	**44.86**	**53.24**	**49.54**	**45.72**	**48.35**	**50.15**	**57.73**	**52.04**	**50.06**	**52.50**

Statement of Authorship

Title of Paper	Learning Motion and Temporal Cues for Unsupervised Video Object Segmentation
Publication Status	□ Published ☑ Submitted for Publication □ Accepted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Submitted to ACM MM 2023

Principal Author

Name of Principal Author (Candidate)	Yunzhi Zhuge		
Contribution to the Paper	Proposed the idea, made partial experiments and wrote the manuscript of the paper.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a s inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	10/05/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Hongyu Gu		
Contribution to the Paper	Make partial experiments		
Signature		Date	12/05/2023

Name of Co-Author	Lu Zhang		
Contribution to the Paper	Discussio		
Signature		Date	12/05/2023

Name of Co-Author	Jinqing Qi		
Contribution to the Paper	Discussion		
Signature		Date	12/05/2023

Name of Co-Author	Huchuan Lu		
Contribution to the Paper	Discussion		
Signature		Date	12/05/2023

Chapter 5

Learning Motion and Temporal Cues for Unsupervised Video Object Segmentation

5.1 Introduction

In this study, we tackle the difficulties associated with unsupervised video object segmentation (UVOS) by introducing an efficient algorithm called MTNet, which simultaneously leverages motion and temporal information. In contrast to previous approaches that focus solely on appearance and motion or temporal relationships, our method unifies these elements within a single framework. MTNet effectively combines appearance and motion features during the encoding process, resulting in a more comprehensive representation.

To effectively utilize the complex long-range context and information present in videos, we incorporate a temporal transformer module that enables efficient inter-frame interactions throughout video clips. Subsequently, a set of decoders is arranged in a cascading manner across all feature levels, aiming to fully exploit their capabilities in generating increasingly accurate segmentation masks.

By doing so, MTNet offers a powerful and streamlined framework that investigates both temporal and cross-modality knowledge to reliably identify and track the primary object with accuracy across a variety of challenging situations. Experiments conducted on an extensive range of datasets underscore the efficacy of our proposed methodology.

5.2 Background

Video Object Segmentation (VOS) is a fundamental task in computer vision, which involves precisely locating and segmenting objects in all frames of a video. Different from semi-supervised video object segmentation (SVOS) which relies on ground-truth masks in the first frame to perform tracking and segmentation in subsequent frames, unsupervised video object segmentation (UVOS) aims to adaptively segment objects without any human intervention. Due to its flexibility, UVOS has broad applications, such as in video editing (Xu et al., 2019; Zeng, Fu, and Chao, 2020), virtual reality (Zhou et al., 2019), and autonomous driving (Hu et al., 2023). However, obtaining precise segmentation results in complex scenes still remains a challenging problem for UVOS.

As a video dense prediction task, some methods naturally associate the near frames via dynamic attention mechanism (Wang et al., 2019c), graph neural networks (Wang et al., 2019b) and pyramid constrained self-attention (Gu et al., 2020). However, these video-based approaches may still struggle to precisely locate the primary object due to several factors. One primary issue is the lack of motion information, which is critical for providing prior knowledge to distinguish the importance and determining the trajectory of objects in unsupervised video object segmentation (UVOS). As a result, high-interference objects can significantly impair the performance of existing algorithms when dealing with complex videos. Moreover, while some methods incorporate frame-to-frame relationships to some extent, their performance may still be inadequate for extremely long-term videos.

Currently, motion-appearance modeling serves as the basis for most advanced unsupervised video object segmentation (UVOS) methods. Optical flow is utilized as a motion guide, whereas appearance cues are extracted from the original images in these approaches. By combining these two modalities, the model can capture the characteristics of the primary object, as well as its movements throughout the video. Although such complementary manner has been proven to be effective in UVOS across several datasets, there is still much room to improve. Firstly, those fusion mechanisms typically depend on sophisticated operations to reach superior performance, and the increasing number of model parameters poses a challenge to satisfy the practical application requirements on devices. Some methods (Zhou et al., 2020b; Zhen et al., 2020; Ji et al., 2021; Yang et al., 2021b) even require a Conditional Random Fields (CRF) (Krähenbühl and Koltun, 2011) post-processing step to obtain well-defined boundaries, which further increases the computational burden of already resource-intensive algorithms. Secondly, they do not explicitly model the video information, which is crucial for tracking the primary object in situations where rapid object displacement and occlusions occur.

After observing the limitations of the methods mentioned above, it is natural to make the suspicion: *what makes for an UVOS algorithm capable of handling various challenging scenarios efficiently?* Generally, an ideal tracker should be able to both i) precisely locate the most important object from a video sequence, 2) robust to handle intricate and long videos in which disappear and occlusion of objects may emerge frequently, and 3) fast to run on edge devices.

Motivated by the aforementioned principles, we present a methodical and efficient algorithm, referred to as MTNet, which *concurrently exploits motion and temporal cues* to address the complex unsupervised video object segmentation task. Different from previous method that merely inherent appearance with motion (in Figure) or modeling temporal relations, our method takes both their merits by collaborating them in a unified framework. More specifically, the Bi-modal Fusion Module is devised to effectively integrate and merge appearance features and motion features during the feature extraction process within encoders, promoting a more complementary representation. In order to comprehensively apprehend the intricate long-range contextual dynamics and information embedded within videos, a Mixed Temporal Transformer is subsequently proposed, which facilitates the achievement of efficacious inter-frame interactions throughout a video clip. the Cascaded Transformer Decoders are incorporated across features of various levels, with the goal of generating progressively more accurate segmentation masks. Taken all those together, MTNet provides an end-to-end, more powerful and compact framework that explores both temporal and cross-modality knowledge to robustly localize and track the primary mask accurately and robustly.

We conduct experiments on a wide range of unsupervised video object segmentation (UVOS) datasets and video salient object detection (VSOD) datasets, making several noteworthy observations:

- MTNet exhibits stronger capability in handling long-term and motion occlusions, positioning itself as a more practical UVOS tool for both academia and industry.
- MTNet achieves state-of-the-art results across a wide range of UVOS and VSOD benchmarks, demonstrating its ability to handle various scenarios and tasks.
- MTNet offers a significant advantage in terms of computational efficiency, as evidenced by its lightweight architecture. Specifically, the compact MTNet achieves a speed of 43.4 frames per second (fps) with a 2080Ti GPU, indicating its potential for real-time applications and suitability for resource-constrained environments.

5.3 Our approach

In this section, we start by presenting a overview of our proposed MTNet architecture in § 5.3.1. Subsequently, we delve into the details of Bi-modal Fusion Module, Mixed Temporal Transformer, and Cascaded Transformer Decoder in § 5.3.2, § 5.3.3, and § 5.3.4, respectively. Lastly, we furnish the details of loss functions in § 5.3.5.

5.3.1 Overview

Given an input video $\mathcal{V} = \{V_i \in \mathbb{R}^{w \times h \times 3}\}_{i=1}^{N}$, the objective of UVOS is to compute binary segmentation masks for the corresponding frames: $\mathcal{S} = \{S_i \in 0, 1^{w \times h}\}_{i=1}^{N}$. To achieve this, we divide the input video and its corresponding length into C clips,

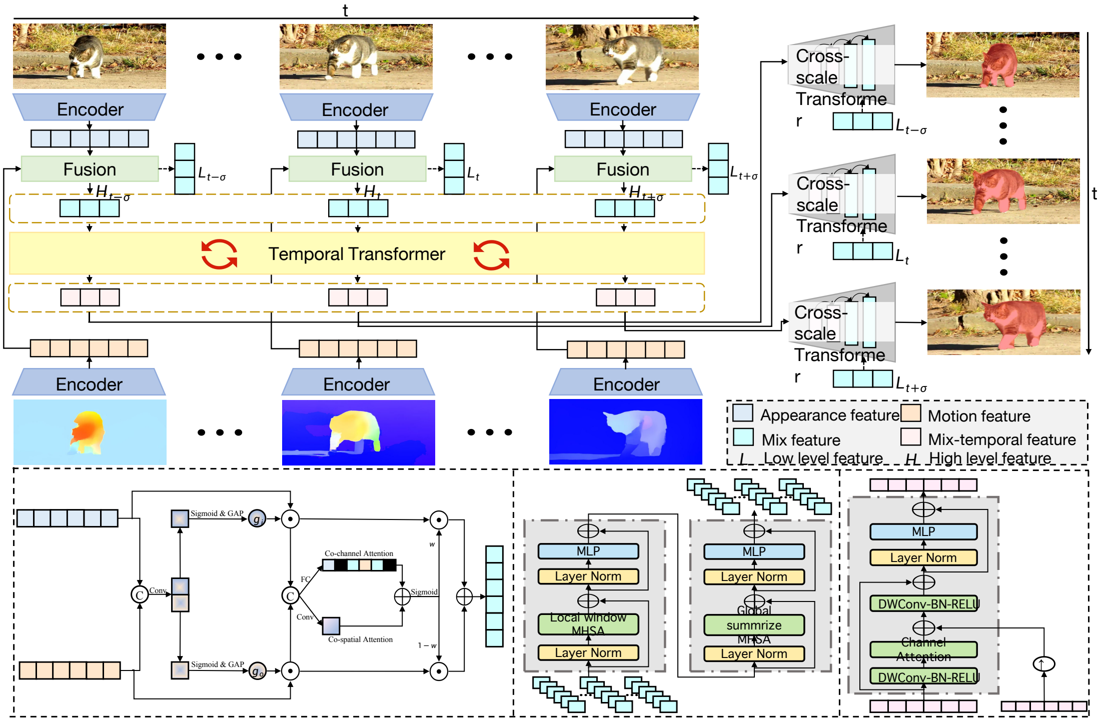


Figure 5.1. (a) The proposed MTNet pipeline utilizes t frames of images and flow maps as input to extract multi-level features. These features at each level are fused by the (b) Bi-modal Feature Fusion Module. Subsequently, the temporal modeling of high-level features are achieved through the (c) Mixed Temporal Transformer. Finally, the output masks are generated using the (d) Dilated Transformer Decoder.

where the number of clips $C = \frac{N}{T}$, and T represents the length of each clip. Following the approach of HFAN (Pei et al., 2022), we employ RAFT (Teed and Deng, 2020) to extract the optical flow, denoted as $\mathcal{O} = \{O_i \in \mathbb{F}^{w \times h \times 3}\}_{i=1}^{N}$. Subsequently, the extracted optical flow is partitioned into C clips, in accordance with the division of the input video. The overall pipeline of MTNet is shown in Figure 5.1(a), which primarily consists of three components: Bi-modal Fusion Module, Mixed Temporal Transformer, and Dilated Transformer Decoder. Initially, we use ConvNeXt (Liu et al., 2022f) as the shared encoder in our approach to extract appearance and motion features from video frame clips and their associated flow maps, respectively. Practically, the extraction of appearance and motion features contains four stages, and we use $k \in [1, 2, 3, 4]$ to denote the each stage. In all both stages of the encoder, Bi-modal Fusion Module is employed to fuse the corresponding appearance and motion features. At the last two stages where the resolutions are reduced to $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, Mixed Temporal Transformer is utilized to model temporal relationships between frames efficiently. Finally, the acquired multi-level features are input into the Dilated Transformer Decoder to generate precise mask predictions for the video clip.

5.3.2 Bi-modal Fusion Module

In various video tasks (Zhu et al., 2017; Xue et al., 2019), optical flow plays a vital role by supplying motion data between successive frames, facilitating the model to learn precise estimations of temporal variations. Prior research in unsupervised video object segmentation (UVOS) (Zhou et al., 2020b; Yang et al., 2021b; Ji et al., 2021; Zhang et al., 2021b; Pei et al., 2022) has employed elaborate designs to align optical flow with video frames, guiding the prediction process. While these approaches have led to significant performance improvements, the intricate operations involved may potentially impede the efficiency of both training and inference stages. In light of this, we introduce a streamlined approach by developing Bi-modal Fusion Modules (BFMs) to efficiently combine features at each level derived from the encoder. We denote the extracted appearance features and motion features as $\{\mathcal{A}_k\}_{k=1}^{K}$ and $\{\mathcal{M}_k\}_{k=1}^{K}$ respectively. For brevity, we use the k-level as an illustrative example, while noting that the operation process can be readily generalized to other levels. As shown in Figure 5.1, The appearance and motion features are initially compressed using two separate 3x3 convolutional layers, followed by a concatenation of the resulting outputs. Subsequently, these combined features are processed through a series of operations designed to obtain the weighted vectors for each modality:

$$\mathcal{F}_k = Conv(Cat(Conv_S(\mathcal{A}_k), Conv_S(\mathcal{M}_k))), \tag{5.1}$$

$$\mathcal{F}_k^A, \mathcal{F}_k^M = Split(\mathcal{F}_k) \tag{5.2}$$

$$g^A = GAP(\sigma(\mathcal{F}_k^A)), \tag{5.3}$$

$$g^M = GAP(\sigma(\mathcal{F}_k^M)), \tag{5.4}$$

where $Conv_S$ and σ denote the 1×1 convolution to shrink feature dimensions and the Sigmoid function. The fused feature $\mathcal{F}_k$ is obtained by concatenating appearance features $\mathcal{A}_k$ and motion features $\mathcal{M}_k$, followed by a 1×1 convolutional layer for initial fusion. The resulting feature is then divided into two groups, $\mathcal{F}_k^A$ and $\mathcal{F}_k^M$, which are subsequently processed by Sigmoid and Global Average Pooling(GAP) operations to derive the weighted vector for each respective modality. Then, the dot production $\odot$ is calculated between input features and the corresponding weighted vector to adaptively enrich the feature of both modalities:

$$\hat{\mathcal{A}}_k = g^A \odot \mathcal{A}_k, \tag{5.5}$$

$$\hat{\mathcal{M}}_k = g^M \odot \mathcal{M}_k, \tag{5.6}$$

The fused features undergo further concatenation and are processed through a parallel of attention operations. Subsequently, a Sigmoid function is applied to constrain the

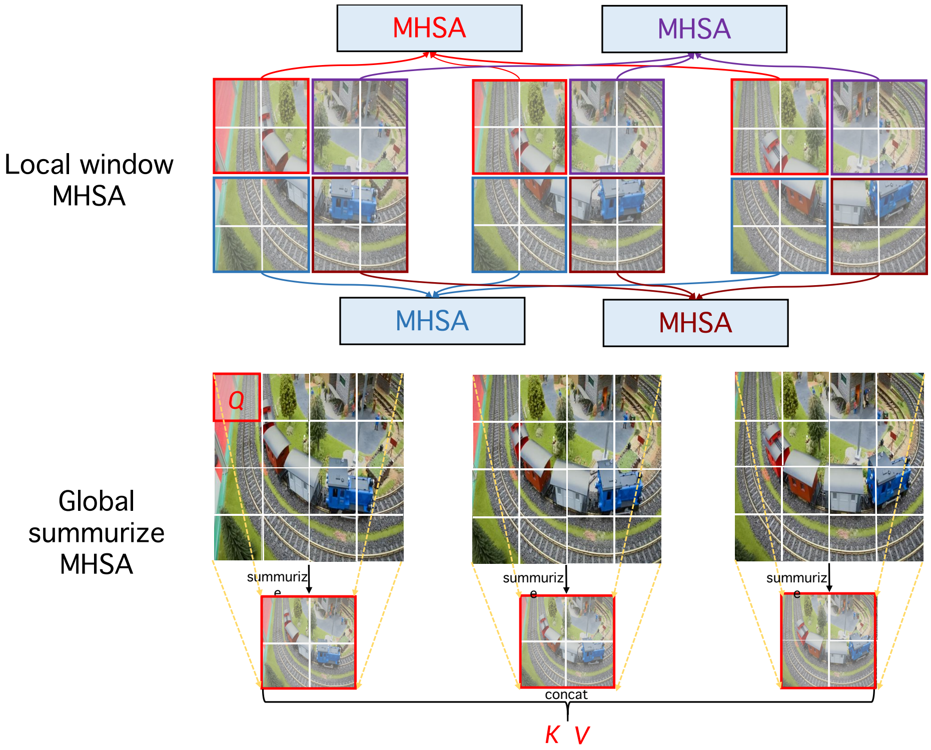


FIGURE 5.2. Illustration of Local Window MHSA (top) and Global MHSA (bottom).

resulting values to the interval [0,1], enabling effective re-weighting:

$$\mathcal{R}_k = Cat(\hat{\mathcal{A}}_k, \hat{\mathcal{M}}_k) \tag{5.7}$$

$$\hat{\mathcal{R}}_k = \sigma(C_{attn}(\mathcal{R}_k) + S_{attn}(\mathcal{R}_k)) \tag{5.8}$$

where the C_{attn} and S_{attn} denotes co-channel attention and co-spatial attention, respectively. With the re-weighted $\hat{\mathcal{R}}_k$, we can finally obtain the output of BFM via:

$$B_k = \hat{\mathcal{R}}_k \odot \hat{\mathcal{A}}_k + (1 - \hat{\mathcal{R}}_k) \odot \hat{\mathcal{A}}_k \tag{5.9}$$

where $k \in \{1 : K\}$ represents the features from different stages of the backbone, in alignment with prior works (Ji et al., 2021; Pei et al., 2022), we set $K = 4$ for our experiments.

5.3.3 Mixed Temporal Transformer

To enhance the temporal relationship, we introduce the Mixed Temporal Transformer, which comprises two transformer layers: the local temporal transformer layer (LTTL) and the global temporal transformer layer (GTTL). This combination effectively captures long-term dependencies in an efficient manner. As is shown in Figure 5.2(d), Given the input feature $B_k \in \mathbf{R}^{T \times H \times W \times d}$, where T, H, W, and d represent temporal length, height, width, and dimension, respectively, local temporal transformer layer (LTTL) divides the input feature into $\frac{H \times W}{M^2}$ windows (Liu et al., 2021b), with each window having the shape $T \times W \times W \times d$. Within each local window, the standard multi-head self-attention is computed. To further enhance the dependencies captured by the LTTL, a global temporal transformer layer (GTTL) is incorporated, which reduces the resolution of keys and values to achieve computational efficiency when calculating self-attention. Specifically, the computational process of the Mixed Temporal Transformer can be represented as:

$$L_k = LTTL(LN(B_k)) + B_k, \tag{5.10}$$

$$L_k^{'} = FFN(LN(L_k)) + L_k, \tag{5.11}$$

$$G_k = GTTL(LN(L_k^{'})) + L_k^{'}, \tag{5.12}$$

$$G_k^{'} = FFN(LN(G_k)) + G_k, \tag{5.13}$$

where LN and FFN represents the layer norm and feed forward network in Transformer (Vaswani et al., 2017). Additionally, we present a comparison between Local Window Multi-Head Self-Attention (LW-MHSA) and Global Summarization Multi-Head Self-Attention (GS-MHSA) in Figure 5.2. This comparison demonstrates that LW-MHSA primarily emphasizes interactions within temporal grids, whereas GS-MHSA effectively captures a more holistic understanding of the integral information. By integrating both transformer layers, the Mixed Temporal Transformer adeptly models the relationships between adjacent frames, which is essential for ensuring consistent object localization within the video clip.

5.3.4 Cascaded Transformer Decoder

Given the input features $\{F_k\}_{k=1}^{K}$ from four stages, where $k \in \{1, 2\}$ from Bi-modal Fusion Module, and $k \in \{3, 4\}$ from Mixed Temporal Transformer, the Cascaded Transformer Decoders (CTDs) can regulate the feature and pass provital information from deep to shallow. Figure 5.1(a) and show the whole decoding process. Specifically, the the whole process in our decoder can be formulated as:

$$\hat{F_k} = \begin{cases} CTD(F_k), & \text{if } k = 4 \\ CTD(F_k, \hat{F_{k+1}}), & \text{if } k = 1, 2, 3 \end{cases} \tag{5.14}$$

Figure5.1(d) provides a detailed illustration of each Cascaded Transformer Decoder, which draws inspiration from recent advancements in incorporating convolutions into transformer architectures(Guo et al., 2022; Liu et al., 2022f). Specifically, we denote the shallower input feature with higher resolution as F_{shal} and the deeper input feature with lower resolution as F_{deep}. The Cascaded Transformer Decoder initially extracts information from F_{shal} as follows:

$$\hat{F}_{shal} = CA(DWConv(F_{shal})), \tag{5.15}$$

where DWConv signifies the sequential operations of a depth-wise convolution, followed by batch normalization and a ReLU activation function, while CA denotes channel attention. Subsequently, the deep feature information is incorporated as follows:

$$F = DWConv(Up(F_{deep}) + \hat{F}shal) + Fshal, \tag{5.16}$$

$$\hat{F} = FFN(LN(F)) + F \tag{5.17}$$

where Up refers to bilinear up-sampling, employed to align the resolution of features, and $\hat{F}$ represents the output. Upon completion of the overall process outlined in Eq.5.14, a 1×1 convolutional layer and bilinear interpolation operation are utilized as the mask decoder to predict the predictions masks P.

5.3.5 Loss Functions

We adopt deep supervision strategies to supervise the predictions at different levels for stabilizing the training process. Specifically, the predictions at frame t can be denoted as $P^t \in \{P^t_s\}^4_{s=2}$, where P^t_1 is the final prediction, and $\{P^t_s\}^4_{s=2}$ represents the auxiliary predictions from various features. Binary cross-entropy loss is used to supervise the training process by comparing P^t and the ground-truth G:

$$\begin{aligned} \mathcal{L} = \frac{1}{H \times W} (\sum_{x,y} \mathcal{L}_{BCE}(P^t_1(x,y), G^t(x,y)) \\ + \lambda \sum_{s=2}^{4} \sum_{x,y} \mathcal{L}_{BCE}(P^t_s(x,y), G^t(x,y)), \end{aligned} \tag{5.18}$$

where (x, y) represents the spatial coordinates within frame t, while λ is set to 0.5 to balance loss terms. During inference and evaluation, we utilize the original prediction results for VSOD, while the `argmax` is further employed to generate binary masks for UVOS.

5.4 experiments

5.4.1 Experimental Setup

UVOS Datasets. We assess our method on four publicly available datasets: DAVIS-16 (Perazzi et al., 2016), FBMS (Ochs, Malik, and Brox, 2013), YouTube-Objects (Prest et al., 2012), and Long-Videos (Liang et al., 2020). DAVIS-16 (Perazzi et al., 2016) comprises 50 high-quality videos with dense annotations, including 30 training videos and 20 validation videos. FBMS (Ochs, Malik, and Brox, 2013) contains 59 natural videos featuring multiple target foreground objects, with 29 sequences designated for training and 30 for testing. YouTube-Objects (Prest et al., 2012) consists of 126 videos spanning 10 object categories, with ground-truth annotations provided sparsely every ten frames. Long-Videos (Liang et al., 2020) encompasses three long videos, totaling over 7,000 frames.

VSOD Datasets. We conduct experiments on four widely-used datasets: DAVIS-16 (Perazzi et al., 2016), ViSal (Wang, Shen, and Shao, 2015) and SegTrack-V2 (Li et al., 2013) and DAVSOD (Fan et al., 2019). Among them, DAVIS-16 (Perazzi et al., 2016) is the same as that employed for UVOS. ViSal (Wang, Shen, and Shao, 2015) and SegTrack-V2 (Li et al., 2013) are earlier datasets for video object segmentation, comprising 17 and 13 video sequences respectively. DAVSOD (Fan et al., 2019) represents a more challenging dataset for video segmentation, characterized by its complex scenes, salience shifts, and diverse attributes,

Evaluation metrics. In accordance with (Ji et al., 2021; Pei et al., 2022), we report mean region similarity ($\mathcal{J}$) and mean boundary accuracy ($\mathcal{F}$) for evaluating UVOS performance. For VSOD, we employ four standard metrics: structure-measure (S_α, α=0.5), maximum enhanced alignment measure (E_ξ^{max}), maximum F-measure (F_β^{max}, β^2=0.3), and mean absolute error (MAE).

Training details. All experiments are conducted using the PyTorch Toolkit. Following HFAN (Zhou et al., 2020b; Pei et al., 2022), we divide the training procedure into two stages: pre-training on the YouTube-VOS (Xu et al., 2018) dataset, followed by fine-tuning on the DAVIS-16 (Perazzi et al., 2016) training set. We employ the tiny version of ConvNext (Liu et al., 2022f) as the shared encoder for extracting both appearance and motion features. RAFT is utilized to generate optical flow maps, which are then converted to a three-channel format. During training, we sample three frames from the same video following the schedule in STCN (Cheng, Tai, and Tang, 2021b) to create a video clip. These video clips undergo data augmentation, including random flips, random crops, random rotations between [-15, 15] degrees, and color jittering; the video order is reversed with a 0.5 probability. All videos are resized to 512 × 512. We employ the AdamW optimizer and Binary Cross-Entropy (BCE) loss for both training stages, while Automatic Mixed Precision (AMP) (Micikevicius et al., 2017) is utilized to expedite the process.

Inference. Upon finishing the training process, we directly evaluate our model on various datasets without applying any dataset-specific fine-tuning.

Our model is specifically designed for offline UVOS, enabling inferences with arbitrary lengths and yielding performance that varies depending on the length. Specifically, for a test video $\mathcal{V}$ and its corresponding flow maps $\mathcal{O}$, both containing N frames, we first partition $\mathcal{V}$ and $\mathcal{O}$ into C clips, where $C = \lfloor \frac{N}{T} \rfloor$ and T represents the test length of each clip. Subsequently, we feed each clip into our model and directly obtain the clip-level results, enabling simultaneous and seamless primary object segmentation and tracking.

5.4.2 Comparisons with State-of-the-Art Models

We show the performance comparisons of our MTNet with other state-of-the-art methods on four UVOS benchmarks and four VSOD benchmarks.

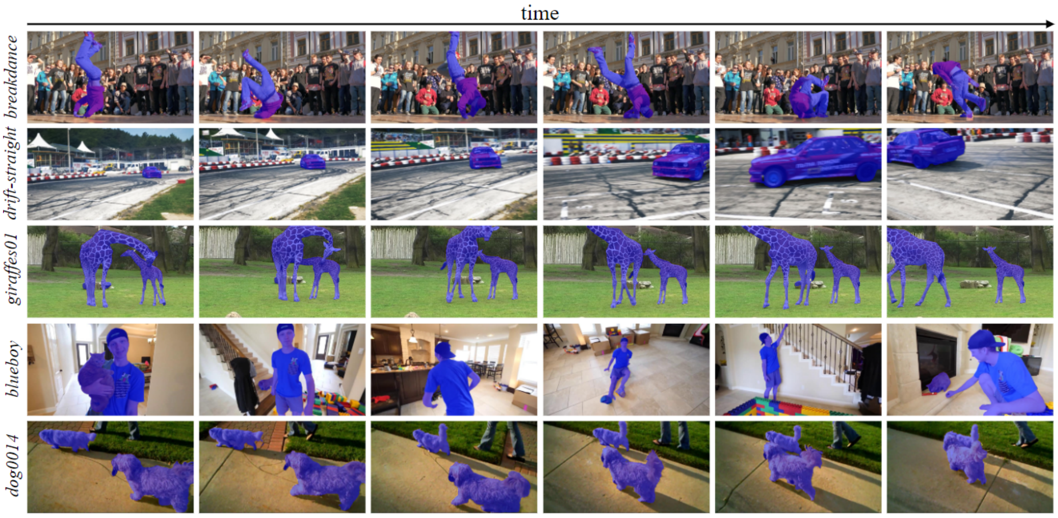
TABLE 5.1. **Quantitative comparison on DAVIS-16 dataset (Perazzi et al., 2016).** ↑ (↓) denotes that the higher (lower) is better. We use the mean region similarity ($\mathcal{J}$), mean boundary accuracy ($\mathcal{F}$) and $\mathcal{J}\&\mathcal{F}$ mean as evaluation metrics. 'PP' denotes post-processing. The top two scores are marked with **Bold** and Underline respectively.

PP	Public.	Method	J			F			J&F	FPS
			Mean ↑	*Recall* ↑	*Decay* ↓	*Mean* ↑	*Recall* ↑	*Decay* ↓	*Mean* ↑	
	$ICCV_{19}$ (Wang et al., 2019b)	AGNN	80.7	94.0	**0.0**	79.1	90.5	**0.0**	79.9	1.9
	$CVRP_{19}$ (Lu et al., 2019)	COSNet	80.5	93.1	4.4	79.5	89.5	5.0	80.0	2.2
	$ICCV_{19}$ (Yang et al., 2019)	AnDiff	81.7	90.9	$\underline{2.2}$	80.5	85.1	$\underline{0.6}$	81.1	2.8
	$AAAI_{20}$ (Gu et al., 2020)	PCSA	78.1	90.0	4.4	78.5	88.1	4.1	78.3	110
✓	$AAAI_{20}$ (Zhou et al., 2020b)	MATNet	82.4	94.5	3.8	80.7	90.2	4.5	81.5	1.3
	$ECCV_{20}$ (Zhen et al., 2020)	DFNet	83.4	94.4	4.2	8.8	89.0	3.7	82.6	3.6
	$AAAI_{21}$ (Liu et al., 2021a)	F2Net	83.1	95.7	**0.0**	84.4	92.3	0.8	83.7	10.0
✓	$CVPR_{21}$ (Ren et al., 2021)	RTNet	85.6	96.1	-	84.7	93.8	-	85.2	-
✓	$ICCV_{21}$ (Ji et al., 2021)	FSNet	83.4	94.5	3.2	83.1	90.2	2.6	84.6	12.5
✓	$ICCV_{21}$ (Yang et al., 2021b)	AMCNet	84.5	$\underline{96.4}$	2.8	84.6	93.8	2.5	84.6	-
✓	$ICCV_{21}$ (Zhang et al., 2021b)	TransportNet	84.5	-	-	85.0	-	-	84.8	3.6
✓	$ECCV_{22}$ (Pei et al., 2022)	HFAN	$\underline{86.8}$	96.1	4.3	$\underline{88.2}$	$\underline{95.3}$	1.1	$\underline{87.5}$	14.4
✓	**Ours**	**MTNet**	**88.7**	**96.9**	4.2	**90.7**	**96.0**	2.9	**89.7**	

Table 5.2. **Quantitative comparison on YouTube-Objects (Prest et al., 2012).** We compare the $\mathcal{J}$ Mean for each category and the average score across categories, with the top two performances denoted by **bold** and underline, respectively.

Method	*Aeroplane*	*Bird*	*Boat*	*Car*	*Cat*	*Cow*	*Dog*	*Horse*	*Motorbike*	*Train*	*Average*
FST (Papazoglou and Ferrari, 2013)	70.9	70.6	57.8	33.9	30.5	41.8	36.8	44.3	48.9	39.2	46.2
LVO (Tokmakov, Alahari, and Schmid, 2017)	86.2	81.0	68.5	69.3	58.8	68.5	61.7	53.9	60.8	66.3	67.2
PDB (Song et al., 2018)	78.0	80.0	58.9	76.5	63.0	64.1	70.1	67.6	58.4	35.3	65.4
AGS (Wang et al., 2019c)	**87.7**	76.7	72.2	78.6	69.2	64.6	73.3	64.4	62.1	48.2	69.7
COSNet (Lu et al., 2019)	81.1	75.7	71.3	77.6	66.5	69.8	76.8	67.4	67.7	46.8	70.5
AGNN (Wang et al., 2019b)	81.1	75.9	70.7	78.1	67.9	69.7	77.4	67.3	**68.3**	47.8	70.8
MATNet (Zhou et al., 2020b)	72.9	77.5	66.9	79.0	73.7	67.4	75.9	63.2	62.6	51.0	69.0
RTNet (Ren et al., 2021)	84.1	80.2	70.0	79.5	71.8	70.1	71.3	65.1	64.6	53.3	71.0
AMC-Net (Yang et al., 2021b)	78.9	80.9	67.4	82.0	69.0	69.6	75.8	63.0	63.4	57.8	71.1
HFAN (Pei et al., 2022)	84.7	80.0	**72.0**	76.1	76.0	71.2	76.9	**71.0**	64.3	61.4	73.4
DATA (Cho et al., 2022)	85.9	83.6	68.4	78.4	77.2	68.3	78.0	70.0	59.4	64.3	73.4
MTNet	83.7	**85.5**	63.5	**83.6**	**79.8**	**72.6**	**81.4**	67.3	56.0	**72.3**	**74.6**

FIGURE 5.3. **Visual comparisons** of videos from diverse scenarios.



DAVIS-16. The DAVIS-16 validation set, comprising 20 videos, is the most commonly used benchmark in UVOS. As illustrated in Table 5.1, MTNet surpasses the recent state-of-the-art method HFAN (Pei et al., 2022) in terms of $\mathcal{J}$ Mean, $\mathcal{F}$ Mean, and $\mathcal{J}\&\mathcal{F}$ Mean, while requiring significantly less inference time. Compared to methods (Yang et al., 2021b; Ji et al., 2021; Pei et al., 2022) that rely on Conditional Random Fields (CRF) or Multi-Scale testing(MS), our approach demonstrates advantages in both inference speed and segmentation quality.

YouTube-Objects. Validation experiments on the Youtube-Objects dataset (Prest et al., 2012) do not require external fine-tuning, which serves to validate the model's generalization ability across diverse scenarios. As shown in Table 5.2, our method is outperformed by other approaches in certain categories, including *Aeroplane*, *Boat*, *Horse*, and *Motorbike*. However, our method exhibits superior performance in the remaining categories, as well as in the overall average results, particularly in *Dog* and *Train*, where it surpasses the second-best method by 4.4% and 12.4%, respectively.

Long-Videos. Long-term videos have been proven to be both challenging and crucial in tracking tasks (Dai et al., 2020; Zhang et al., 2021d), warranting increased attention in unsupervised video object segmentation. The Long-Videos dataset (Liang et al., 2020) comprises three video sequences, each averaging 2500 frames. In this benchmark, we compare our method not only with other UVOS approaches but also with more competitive SVOS methods that incorporate additional priors. As shown in Table 5.3, the proposed MTNet surpasses HFAN (Pei et al., 2022) by 1.2% in $\mathcal{J}\&\mathcal{F}$ Mean, highlighting the effectiveness of the temporal modules in handling long-term videos. However, a performance gap still exists when compared to SVOS methods.

Main results on VSOD datasets. We compare the performance of our MTNet in three VSOD datasets, *i.e.*, DAVIS-16 (Perazzi et al., 2016), ViSal (Wang, Shen, and

TABLE 5.3. **Quantitative comparison on Long-Videos dataset.** The best SVOS and UVOS results are marked with underline and bold respectively.

Public.	Method	J			F			J&F
		Mean ↑	*Recall* ↑	*Decay* ↓	*Mean* ↑	*Recall* ↑	*Decay* ↓	*Mean* ↑
Performance of SVOS Methods								
$ICCV_{19}$ (Gu et al., 2020)	STM	78.1	90.0	4.4	78.5	88.1	4.1	78.3
$NIPS_{20}$ (Zhou et al., 2020b)	AFB-URR	82.7	91.7	11.5	83.8	91.7	13.9	83.3
$NIPS_{21}$ (Zhen et al., 2020)	AOT	83.2	-	-	85.4	-	-	84.3
Performance of UVOS Methods								
$AAAI_{20}$ (Gu et al., 2020)	AGNN	68.3	77.2	13.0	68.6	77.2	16.6	68.5
$AAAI_{20}$ (Zhou et al., 2020b)	MATNet	66.4	73.7	10.9	69.3	77.2	10.6	67.9
$ECCV_{20}$ (Zhen et al., 2020)	HFAN	80.2	91.2	**9.4**	83.2	96.5	7.1	81.7
Ours	**MTNet**	**79.6**	**91.2**	9.5	**85.8**	**96.7**	**6.7**	**82.7**

TABLE 5.4. **Quantitative comparison on benchmark VSOD datasets.** ↑ (↓) denotes that the higher (lower) is better. We use the Mean Absolute Error (MAR), max F-measure (F_m), S-measure (S_m), and max E-measure (E_m) as evaluation metrics. Bold and underline denotes the best and secondary results. * indicates that the reported results were obtained through our own re-measurement process.

Public.	Dataset	DAVIS-16 (Perazzi et al., 2016)				ViSal (Wang, Shen, and Shao, 2015)				SegTrack-V2 (Li et al., 2013)				DAVSOD (Fan et al., 2019)			
	Metric	$S_m \uparrow$	$E_m \uparrow$	$F_m \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_m \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_m \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_m \uparrow$	$MAE \downarrow$
$ECCV_{18}$ (Song et al., 2018)	PDB	0.882	0.951*	0.855	0.028	0.907	-	0.888	0.032	0.864	-	0.800	0.024	0.698	-	0.572	0.116
$CVPR_{19}$ (Fan et al., 2019)	SSAV	0.893	0.948	0.861	0.028	0.943	0.977*	0.939	0.020	0.851	0.918*	0.801	0.023	0.755	0.806	0.659	0.084
$ICCV_{19}$ (Yan et al., 2019)	RCR	0.886	0.947	0.848	0.027	0.922	0.955*	0.906	0.026	0.843*	0.878*	0.780*	0.035*	0.741	0.083	0.653	0.087
$AAAI_{20}$ (Gu et al., 2020)	PSCA	0.902	0.961	0.880	0.022	0.946	0.983*	0.940	0.017	0.865	0.907	0.810	0.025	0.741	0.793	0.656	0.086
$ICCV_{21}$ (Zhang et al., 2021c)	DCFNet	0.914	0.969*	0.900	0.016	0.952	0.990*	0.953	**0.010**	0.883	0.935*	0.839	0.015	0.741	0.805*	0.660	0.074
$ICCV_{21}$ (Ji et al., 2021)	FSNet	0.920	0.970	0.907	0.020	0.923*	0.972*	0.907*	0.023*	0.871*	0.910*	0.804*	0.024*	0.773	**0.825**	0.685	0.072
$ECCV_{22}$ (Pei et al., 2022)	HFAN	0.938	0.983	0.935	**0.008**	0.944*	0.980*	0.927*	0.014*	0.876*	0.951*	0.828*	0.015*	0.754*	0.788*	0.645*	0.077*
Ours	**MTNet**	**0.951**	**0.987**	**0.944**	0.010	**0.959**	**0.991**	**0.959**	0.011	**0.899**	**0.962**	**0.875**	**0.013**	**0.780**	**0.825**	**0.688**	**0.067**

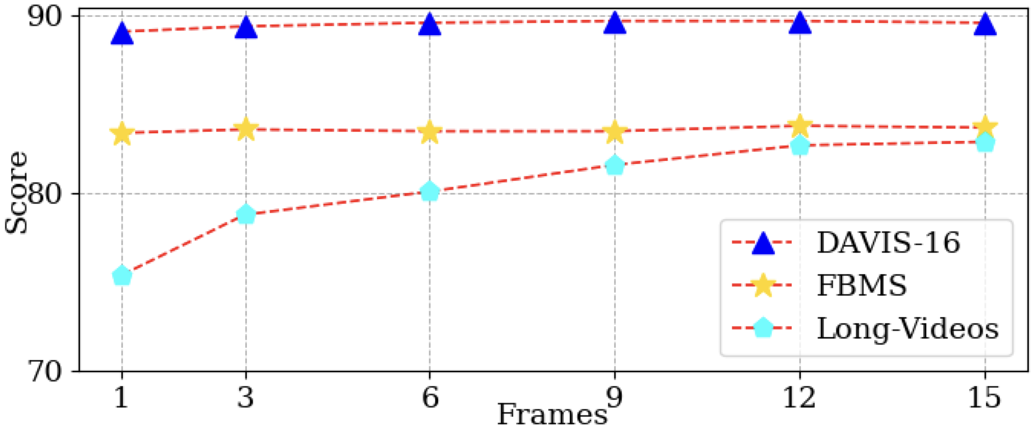


FIGURE 5.4. **Influence of clip length during inference.**

Shao, 2015), SegTrack-V2 (Li et al., 2013) and DAVSOD (Fan et al., 2019). Specifically, we compare MTNet with 10 state-of-the-art VSOD methods. The results are sourced either directly from the original publications or re-measured by us, adhering to a strict evaluation process that employs the original testing codes and model weights in their projects. As shown in Table 5.4, out proposed MTNet exhibits exceptional performance in the majority of datasets, consistently achieving the best results across various evaluation metrics. In Davis-16, MTNet outperforms previous best model HFAN (Pei et al., 2022) by 1.0% in terms of S_m and F_m, and similar leading trend can also be observed in FMBS abd DAVSOD.

Visualization Results. We present qualitative results of our method in Figure 5.3, illustrating the capability of our approach to produce high-quality outcomes. In the 1^{st} row, the person in red shirt and blue jeans is dancing street dance, the pose of which changes continuously. Our MTNet could consistently captures the person despite the background is full of onlookers, which is quite distracting. In the 1^{st} row, we observe an individual wearing a red shirt and blue jeans engaging in a street dance performance, characterized by a continuous change in poses. Despite the presence of numerous onlookers in the background, which could potentially be quite distracting, our method consistently and accurately captures the person of interest. Similar phenomena can be observed in 2^{nd} and 4^{th} rows as well, where the scale and appearance of the racing car undergo dramatic changes, and the 'blueboy' activates freely throughout the room. In the videos where multiple objects co-exists (3^{nd} and 5^{th} rows), our method can still achieve accurate tracking and segmentation. These cases demonstrates the robustness of our approach in effectively distinguishing the target subject from complex and dynamic backgrounds.

5.4.3 Ablation Studies

In this section, we conduct ablation studies on the DAVIS-16 and FBMS datasets. We use the ConvNext version of MTNet as the standard model in our ablation study, with the results reported in Table 5.5 (#1).

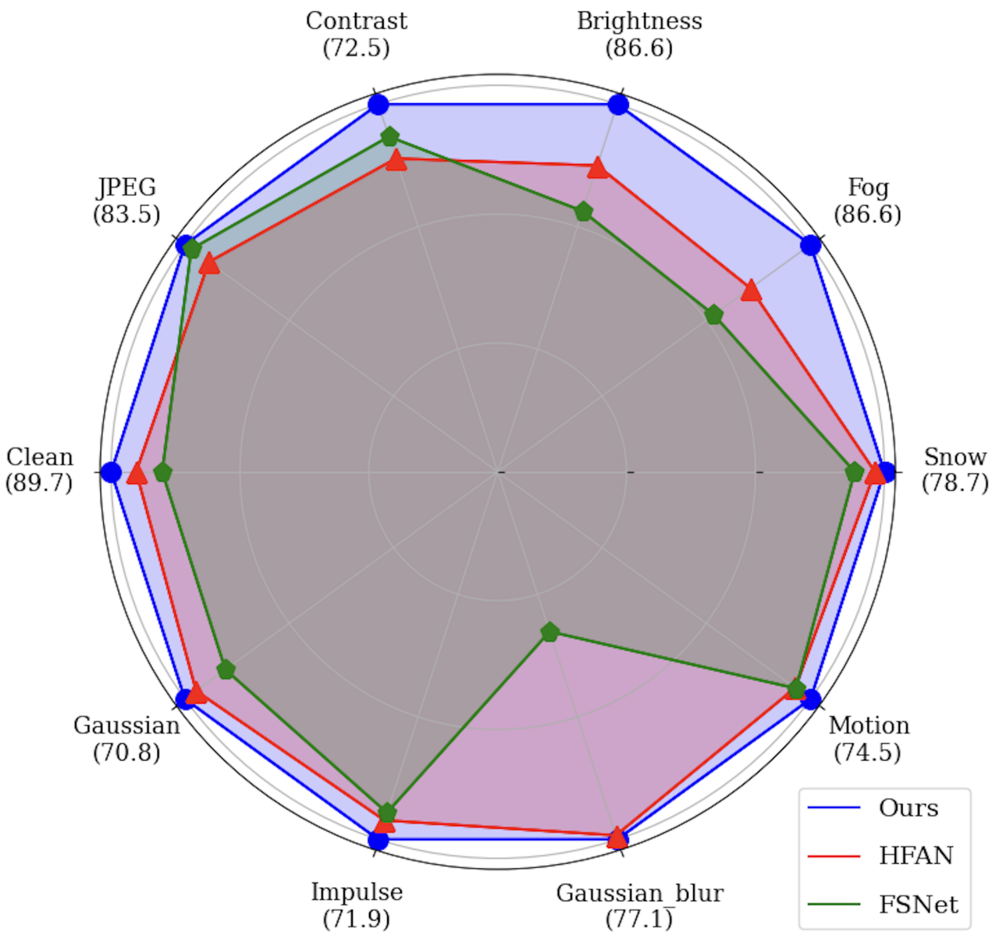


FIGURE 5.5. $\mathcal{J}\&\mathcal{F}$ **Mean when applying various corruptions** on DAVIS-16 (Perazzi et al., 2016).

Design choices of modules. We introduce the Bi-modal Fusion Module (BMF), Mixed Temporal Transformer (MTT), and Cascaded Transformer Decoder (CTD) to perform robust and accurate object tracking and segmentation. We validate the effectiveness of each component. Table 5.5 (#2) presents the results of our baseline, which is a simple combination of the encoder and FPN without any specific designs. Experiments #3-#7 add different components to the baseline in a controlled manner. Compared with the baseline results, we can observe that both components contribute to the performance. In particular, adding either MTT or CTD results in a noticeable gain, with both components increasing the *average* $\mathcal{J}\&\mathcal{F}$ Mean by 3.6. When both components are used, the absolute increase further reaches 5.5, yielding the best performance among all configurations.

Impact of modality. To investigate the influence of input modality on performance, we conduct experiments and present the results in Table 5.5(8-11). In these experiments, we employ a single-stream encoder for feature extraction, and the Bi-modal Fusion Module is omitted. The terms 'Add' and 'Concat' indicate that images and flow maps are pre-fused using addition or concatenation before being fed into the model. The observed decline in performance for both modifications underscores the importance of utilizing both modalities and the dual-branch architecture.

Influence of training stages. Since our method undergoes a two-stage training process, *i.e.*, pre-training on YouTube-VOS (Xu et al., 2018) first and then fine-tuning on DAVIS-16 (Perazzi et al., 2016), we conduct ablation experiments to evaluate the impact of each stage. The results can be seen in Table 5.5(#12,#13). When removing the fine-tuning stage and pre-training stage, the model experiences significant performance decreases of 9.9 and 6.8, respectively, indicating that both training stages are crucial.

Clip length. We examine the impact of varying clip lengths during the inference stage. Our experiments are conducted on three benchmark datasets: DAVIS-16 (Perazzi et al., 2016), FBMS (Ochs, Malik, and Brox, 2013), and Long-Videos (Liang et al., 2020). As illustrated in Figure 5.4, the $\mathcal{F}\&\mathcal{J}$ Mean for DAVIS-16 and FBMS exhibits only minor changes as the clip length increases. Conversely, the Long-Videos dataset demonstrates a significant improvement in performance as a function of clip length, with this trend persisting until saturation is reached as t extends. Based on these findings, we set $t = 12$ as the standard clip length in our experiments.

Robustness to corruptions. Robustness is a crucial aspect in various domains, including segmentation (Xie et al., 2021a) and autonomous driving (Xie et al., 2023; Ge et al., 2023). We sample nine common corruptions from ImageNet-C (Hendrycks and Dietterich, 2019) and apply the most intense degrees to the DAVIS-16 (Perazzi et al., 2016) validation set. Importantly, all results are obtained through zero-shot testing, without any fine-tuning. As depicted in Figure 5.5, our method consistently exhibits superior robustness compared to other approaches under a diverse range of corruptions. This finding underscores the potential of our method to deliver reliable performance in challenging and unpredictable environments.

Table 5.5. **Ablation study on DAVIS-16** (Perazzi et al., 2016) and FBMS (Ochs, Malik, and Brox, 2013). We use gray, green, pink and yellow colors to denote the proposed method, ablations of designed modules, input modality and training process, respectively. △ denotes the performance change (averaged over benchmarks) compared with the MTNet.

#	Method	DAVIS-16	FBMS	△
1	MTNet	89.7	83.8	-
2	Baseline	86.8	75.9	**-5.5**
3	*w/* BMF	87.6	77.0	**-4.5**
4	*w/* MTT	88.7	81.2	**-1.9**
5	*w/* CTD	88.9	81.0	**-1.9**
6	*w/* BMF+MTT	89.2	80.5	**-2.0**
7	*w/* BMF+CTD	89.3	82.1	**-1.1**
8	*w/o* Appearance	82.0	63.1	**-16.0**
9	*w/o* Motion	85.9	79.2	**-4.3**
10	Add Motion&Appearance	87.1	72.3	**-7.1**
11	Concat Motion&Appearance	86.3	77.2	**-5.0**
12	*w/o* Fine-tuning	82.7	71.2	**-9.9**
13	*w/o* Pre-training	84.8	75.2	**-6.8**

Statement of Authorship

Title of Paper	BEVS4: Semi-supervised Bird's-eye-view Semantic Segmentation	
Publication Status	□ Published ☑ Submitted for Publication	□ Accepted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to ACM MM 2023	

Principal Author

Name of Principal Author (Candidate)	Yunzhi Zhuge		
Contribution to the Paper	Proposed the idea, made partial experiments and wrote the manuscript of the paper.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	10/05/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Jiayuan Zhou		
Contribution to the Paper	Make partial experiments		
Signature		Date	12/05/2023

Name of Co-Author	Lijun Wang		
Contribution to the Paper	Discussion, revision of paper		
Signature		Date	12/05/2023

Name of Co-Author	Yifan Wang		
Contribution to the Paper	Discussion		
Signature		Date	12/05/2023

Name of Co-Author	Huchuan Lu		
Contribution to the Paper	Discussion		
Signature		Date	12/05/2023

Chapter 6

BEV-S^4 Semi-supervised Bird's-eye-view Semantic Segmentation

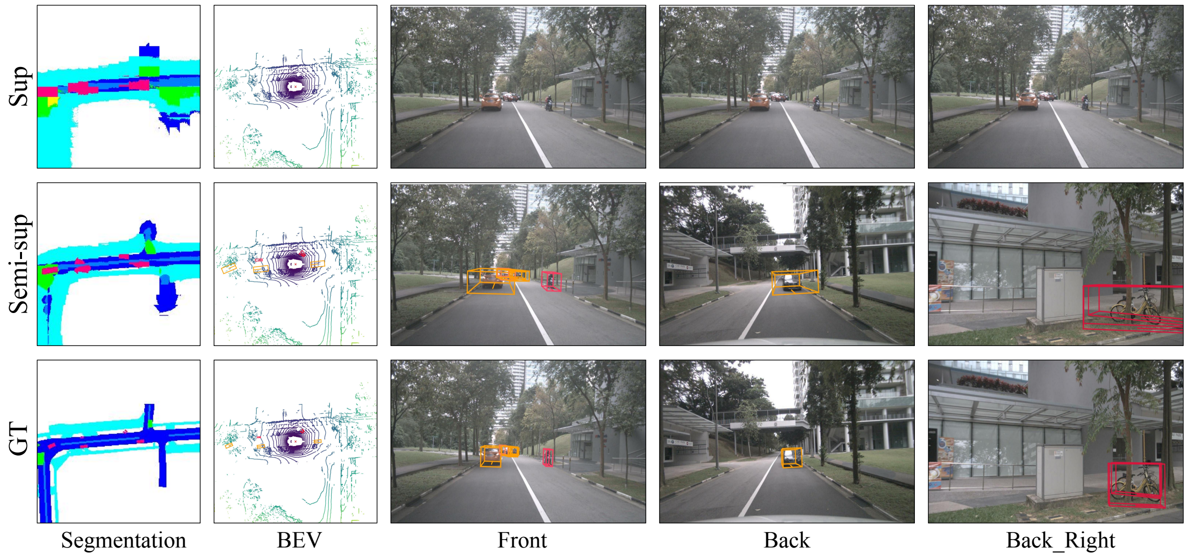


FIGURE 6.1. Visualizations of the supervised baseline and our semi-supervised method using 3D bounding boxes under the BEV view (2nd column) and selected image views (3rd-5th columns) are presented. The 3D bounding boxes for each view are transformed from the predictions of BEV segmentation (1st column). It is important to note that our method is purely vision-based, and the LiDAR maps in both rows represent the ground truth for illustrative purposes.

6.1 Introduction

Recently, bird's-eve-view semantic segmentation have drawn much attention due to the emergence of autonomous driving. However, most of the previous methods are trained using fully-supervised learning paradigm, requiring large amounts of human annotations which are labor-intensive to obtain. Drawing inspirations from the prevalent of semi-supervised learning in 2D semantic segmentation, we present a simple-yet-effective baseline approach utilizing a teacher-student framework for bird's-eye-view semantic segmentation, which achieves comparable results to fully-supervised counterpart with limited annotations. In this work, we aim to learn BEV semantic segmentation with limited manual annotations. To address this novel and challenging task, we propose BEV-S^4: a baseline method for **s**emi-**s**upervised **s**emantic **s**egmentation in **BEV** space. Employing a teacher-student architecture, our method emphasizes BEV pseudo labeling, intricately designed with data perturbations and thresholding. To fully utilize the spatial-temporal properties of BEV, we further propose a teacher temporal ensemble approach to enhance the quality of pseudo-labels without introducing extra computational costs during inference. Extensive experiments on nuScenes dataset (Caesar et al., 2020) demonstrate that our method significantly boosts the performance under various proportions of labeled training data. We believe our method would serve as a strong baseline and attract more attention to learning BEV perception with fewer labels.

6.2 Background

Bird's-eye-view (BEV) perception tasks, such as 3D object detection (Huang et al., 2021), BEV segmentation (Zhou and Krähenbühl, 2022), motion prediction (Zhou and Krähenbühl, 2022), and lane detection (Zhou and Krähenbühl, 2022), play a crucial role in the fields of autonomous driving and robotics. These tasks are essential for enabling vehicles and robots to accurately perceive their surroundings, understand the spatial relationships between objects, and make informed decisions based on the information they own.

Among the various BEV perception tasks, BEV segmentation is of particular interest as it focuses on performing semantic segmentation on the surrounding regions and objects. This encompasses, among others, the categorization of areas such as drivable areas, lanes, car parks, and vehicles. As a fundamental step towards constructing the BEV map, BEV segmentation has been paid significant attention in recent times.

However, the development of high-performance BEV segmentation algorithms necessitates the utilization of expensive and labor-intensive manual annotation processes. Take the widely used nuScenes as an example, the segmentation annotations are obtained by transforming 1.4M 3D bounding boxes annotations over 40K frames to the BEV space. In comparison to the traditional manual annotation processes, which can be both limited in scope and financially costly, the acquisition of raw data directly from driving vehicles offers a scalable and cost-effective alternative. This makes it an attractive point for researchers looking to overcome the challenges associated with manual annotation and advance the algorithms in this field.

The utilization of the plentiful yet unlabeled data for advancing BEV perception constitutes a crucial and prospective direction of research, which has received limited attention until now. As completely getting rid of the annotations is unrealistic, in this work we seek into the paradigm of semi-supervised learning, which has been proved to be effective in a wide range of 2D dense prediction tasks, *e.g.*semantic segmentation, object detection and point cloud semantic segmentation. However, directly migrating the existing semi-supervised pipeline to solve bird's-eye-view can be difficult, as there exists a huge modality gap between bird's-eye-view and image view.

Semi-supervised semantic segmentation methods that have been developed recently primarily rely on consistency training, which shows benefits such as increased stability and better generalization capabilities. A representative work in the field of semi-supervised semantic segmentation is CowMix (French et al., 2019), which implements a consistency enforcement strategy on the outputs generated from mixed inputs with their corresponding predictions using the MixUp (Zhang et al., 2017). Other subsequent studies have either used multiple decoders on unlabeled data to produce various predictions while ensuring consistency between the main decoder's outputs and the others (Ouali, Hudelot, and Tami, 2020), or employed cross pseudo-supervision, in which the consistency between two segmentation networks is ensured by supervising one branch with the pseudo labels from the other and vice versa (Chen et al., 2021). Despite advancements in the field of semi-supervised semantic segmentation, the determinants of its efficacy are yet to be explicitly established due to the insufficiency of evaluation criteria. In making quantitative comparisons, the lack of strict alignment between variables such as segmentation networks and input resolutions can result in unfair comparising results.

Therefore, our motivations towards resolving the issue of semi-supervised bird's-eye-view semantic segmentation are:

- What is the potential of exploring established semi-supervised learning approaches and designs to augment the effectiveness of our task?
- How to address our task by considering its intrinsic essence and seeking more effective solutions?

We start designing from the perspective of data augmentation, the goal of which is to make the model invariant and robust to various augmentations applied to the input, and is a widely explored approach in the consistency training based semi-supervised semantic segmentation methods. Previous works (Chen et al., 2021; Yuan et al., 2021; Yang et al., 2022; Zhao et al., 2022) already propose various data augmentation strategies for generating discrepant inputs and enforcing consistency on them. However, due to the lack of uniform standards, many data augmentations are difficult to quantify in terms of their transferability in other pipelines. CPS (Chen et al., 2021) Incorporates with the Cutmix (Yun et al., 2019) as the data augmentation in which both the input two source images and the two pseudo segmentation maps are mixed for inputs of

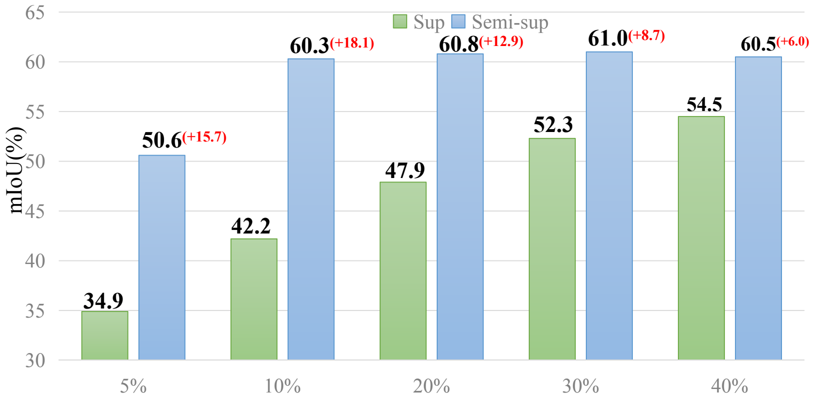


FIGURE 6.2. **mIoU comparisons of supervised baseline and our semi-supervised method on nuScenes.**

networks and cross supervision recpectively. Following RandAugment (Cubuk et al., 2020), (Yuan et al., 2021) establish a pool or operations which is composed of 16 image transformations to form the strong augmentation. ST++ (Yang et al., 2022) further presents a thorough empirical and systematic analysis of the efficacy of strong data augmentation techniques on unlabeled data. The results indicate that performing strong augmentations on labeled data may have a detrimental effect on the integrity of the underlying clean data distribution. In contrast, implementing strong augmentation solely on unlabeled data leads to superior performance outcomes. Inspired by those prior works, we carefully design a set of data augmentation techniques for our task, aims at enhancing the performance and robustness of our model.

Besides, BEV perception is a temporal task, where the implementation of efficient temporal fusion strategies holds great promise for enhancing performance. BEVFormer (Li et al., 2022) incorporates temporal self-attention to recurrently integrate historical BEV information with temporal information, resulting in improved velocity estimation of moving objects and detection of heavily occluded objects with minimal computational overhead. SOLOFusion (Park et al., 2022a) highlights the limitation of utilizing only a limited number of recent frames, which restricts the localization capacity and undermines the advantages of prior works' temporal fusion techniques. After exploiting the utilization of a 16-frame BEV cost volume, which already demonstrates improved results, they make a further step by employing short-term temporal fusion in conjunction with a computationally efficient sampling module to achieve even better performance. Considering that incorporating an additional external temporal module may introduce computational complexity, we present a straightforward design that maintains the original architecture of the network. Our approach leverages a temporal ensemble technique applied to the predictions generated by the teacher network, thereby enhancing the reliability of the generated pseudo-labels.

We adopt a fully-supervised method PETR-v2 as our baseline. The student model learns both the manual annotation on the labeled sequences and also the teacher model's output on the unlabeled sequences. To avoid harmful pseudo-label biases, we carefully design the pseudo labeling strategies in BEV space, including data perturbations and thresholding. We further leverage the spatial-temporal properties of BEV perception to perform stronger semi-supervised learning. The highly structured video sequences provide rich and informative cues, which can be used to construct more accurate training signals on unlabeled data. Specifically, we introduce a teacher temporal ensemble strategy to generate more reliable pseudo labels by aggregating information from multiple historical frames. This approach brings no additional computational overhead to the model inference, as the ensemble operation is performed only during the training phase.

Without bells and whistles, our semi-supervised method yields significant improvements over the supervised baseline (see Figure 6.2) on nuScenes dataset. Moreover, we compare the results of transforming the segmentation maps to 3D bounding boxes in both the BEV view and multiple image views (see Figure 6.1). These comparisons highlight the potential of our semi-supervised pipeline in assisting downstream tasks. Importantly, the performance gains are consistent across various settings, ranging from 5% to 40% labeled data. Remarkably, our method attains comparable performance to the fully supervised counterpart while using less than half of the labels.

In summarize, this chapter presents several substantial contributions to the field, which are meticulously described as follows:

- Observing the potential impact of high annotation costs on the progress of BEV perception tasks, we aim to investigate and implement label-efficient approaches to enhance the efficiency of this task.
- We introduce a simple-yet-effective framework to solve the new problem of semi-supervised semantic segmentation in BEV space, with careful designs on data augmentation strategy and BEV pseudo labeling.
- We propose the teacher temporal fusion to fully utilize historical information for generating more reliable pseudo-labels, without adding extra computational cost during inference.
- We conduct extensive experiments with varied settings and show that our proposed BEV-S^4 can achieve significant improvements when learning with partially labeled data.

6.3 Method

In this work, we aim to solve semi-supervised BEV semantic segmentation. Figure 6.3 overviews the framework of our proposed BEV-$S4$ method. The student model is trained on both labeled and unlabeled data simultaneously via consistency training. The teacher model is built by a temporal ensemble to capture temporal coherence for generating high-quality pseudo labels for unlabeled data. Both student and teacher models adopt PETRv2 (Liu et al., 2022b) for its simplicity, and our framework is also

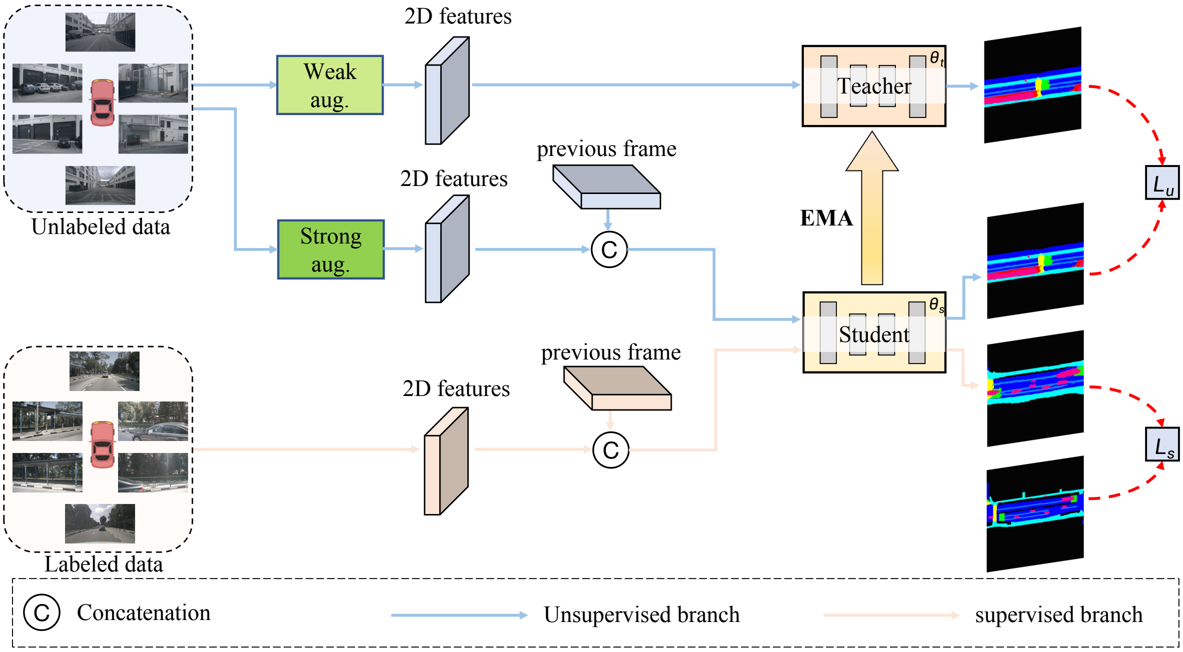


FIGURE 6.3. **Overall framework of BEV-S^4.** BEV-S^4 adopts a teacher-student architecture. The teacher model parameterized by θ_t generates pseudo-labels on unlabeled images, and is slowly progressed via exponential moving average (EMA) of student. The student model parameterized by θ_s is trained by jointly minimizing the supervised loss $\mathcal{L}_s$ on labeled data with ground truth and the supervised loss $\mathcal{L}_u$ on unlabeled data with pseudo-labels.

applicable to other BEV segmentation models. To further improve the generalization ability of semi-supervised learning, a variety of data augmentation strategies are also investigated from the perspective of BEV perception. In the following, we first revisit PETRv2 (Liu et al., 2022b) in Section 6.3.1. Section 6.3.2 and Section 6.3.3 introduce the proposed teacher-student architecture as well as teacher temporal ensemble in detail. We finally elaborate on the proposed data augmentation strategies in Section 6.3.4.

6.3.1 Revisiting PETRv2

The pipeline of PETRv2 contains three key components, *i.e.*, temporal modeling, feature-guided position embedding and BEV map prediction. We briefly review them as follows and refer the readers to (Liu et al., 2022b) for more details.

Temporal Modeling. PETRv2 implements temporal modeling by 3D coordinates alignment (CA) and feature-guided position encoder (FPE) to fuse the information between current and historical input. Denoting camera cooedinate as $c(t)$, lidar coordinate as $l(t)$ and ego coordinate as $e(t)$ at frame t, the coordinates of 3D points from a previous frame $t-1$ are aligned to frame t by using global coordinate space to bridge two frames:

$$P_i^{l(t)}(t-1) = T_{l(t-1)}^{l(t)} P_i^{l(t-1)}(t-1), \tag{6.1}$$

where $P_i^{l(t-1)}(t-1)$ denotes the 3D points projected from i-th camera in frame $t-1$, and the temporal transformation can be calculated by:

$$T_{l(t-1)}^{l(t)} = T_{e(t)}^{l(t)} T_g^{e(t)} {T_g^{e(t-1)}}^{-1} {T_{e(t-1)}^{l(t-1)}}^{-1}, \tag{6.2}$$

The aligned point sets $\left[P_i^{l(t)}(t-1), P_i^{l(t)}(t)\right]$ are further utilized to generate 3D PE.

Feature-guided Position Embedding. Different from the image independent 3D position embedding (3D PE) in PETR, PETRv2 adopts the feature-guided position encoder by:

$$PE_i^3d(t) = \xi(F_i(t)) * \psi(P_i^{l(t)}(t)), \tag{6.3}$$

where $F_i(t)$ denotes the 2D features of the i-th camera. ψ and ξ are two multi-layer perceptions(MLPs) for generating attention weights and 3D PE respectively. The addition of 3D PE with 2D features and the projected 2D features serve as key and value for transformer decoder respectively.

BEV Map Prediction. In PETRv2, seg query is introduced to BEV segmentation and each of which corresponds to a patch region of the BEV map. Initialized with fixed anchor points and projected by a MLP with two linear layers, those seg queries are then input to the transformer decoder and interact with the image features. Finally, a segmentation head is applied to predict the segmentation map with the updated seg queries as inputs.

6.3.2 Teacher-Student Architecture

Consistency regularization methods are predominant in semi-supervised semantic segmentation. Motivated by their great success, we adopt the consistency learning philosophy for semi-supervised BEV semantic segmentation. As depicted in Figure 6.3, our proposed BEV-S^4 consists of a teacher model and a student model, parameterized by θ_t and θ_s, respectively. We use the exponential moving average (EMA) of the student model to gradually evolve the teacher model:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{6.4}$$

where α denotes the momentum parameter, which is set to 0.999 following previous works.

To better leverage the enormous amount of unlabeled data, we train the student model via supervised learning on labeled data regularized by the consistency between student and teacher predictions on unlabeled data. To this end, labeled and unlabeled images are randomly sampled according to a predefined data sampling ratio s_r to form a training data batch in each training iteration. The overall loss function for training the student model is defined as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u, \tag{6.5}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ denote supervised loss on labeled images and unsupervised consistency regularization on unlabeled images, respectively, and λ is the hyper-parameter to balance the loss weights. For labeled images, the weighted CE loss (Liu et al., 2022b) is utilized to supervise the predicted BEV map:

$$\mathcal{L}_s = \frac{1}{N} \sum_{k=0}^{K} \sum_{i=0}^{N} \omega^s \hat{y}_{i,k} \log(y_{i,k}) + (1 - \hat{y}_{i,k}) \log(1 - y_{i,k}), \tag{6.6}$$

where N is the number of pixels, K is the number of object categories, $\hat{y}$ is the ground-truth label, and y is the BEV map predicted by the student model, with $y_{i,k}$ indicating the probability of the i-th pixel belonging to the k-th category. ω^s represents the weight of positive samples and is obtained by calculating the proportion between the negative samples and the positive samples in ground truth. It should be noted that one pixel in the BEV space may belong to multiple semantic categories, therefore, we compute the binary weighted cross-entropy loss for each category independently.

For unlabeled images, we use the teacher model to generate the corresponding pseudo labels for consistency regularization. Although the teacher model is more reliable than the student model, the generated pseudo labels may inevitably contains erroneous predictions. We thus design a confidence mining based label selection method to alleviate the impact of noisy pseudo labels on the training process. The basic idea is to filtering out unreliable pseudo labels with low prediction confidence. To this purpose, we first compute a confidence map ω^u according to the predicted pseudo label $\tilde{y}$:

$$\omega^u_{i,k} = \begin{cases} 1, & \text{if } \tilde{y}_{i,k} > \tau_{up} \text{ or } \tilde{y}_{i,k} < \tau_{low}, \\ 0, & \text{otherwise}, \end{cases} \tag{6.7}$$

where τ_{up} and τ_{low} are pre-defined thresholds. We then implement the unsupervised consistency regularization using the following confidence-aware cross-entropy loss:

$$\mathcal{L}_u = \frac{1}{N} \sum_{k=0}^{K} \sum_{i=0}^{N} \omega^u_{i,k} (\tilde{y}_{i,k} \log(y_{i,k}) + (1 - \tilde{y}_{i,k}) \log(1 - y_{i,t})). \tag{6.8}$$

where unreliable pseudo labels (*i.e.*, those with $\omega^u_{i,k} = 0$) can be effectively ignored.

6.3.3 Teacher Temporal Ensemble

Recent study (Liu et al., 2022b; Li et al., 2022; Park et al., 2022b) indicates that temporal feature learning using historical information can significantly benefit BEV perception. To justify this idea under the semi-supervised learning framework, we design a teacher ensemble approach (See Figure 6.4) to capture temporal consistency in BEV data. Specifically, given the BEV features $\{B_h | h = 1, 2, \ldots, T\}$ of T historical

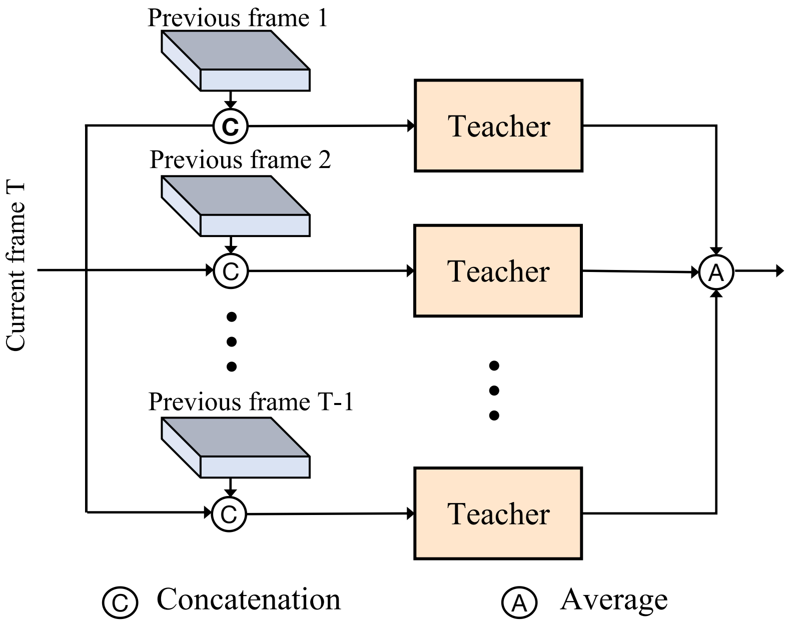


FIGURE 6.4. **Teacher temporal ensemble.** Features of multiple previous frames are aligned and fused with current features respectively. Averaging aggregation is then applied on prediction maps to obtain more reliable pseudo-labels.

inputs, we first warp these feature into the current view following the procedures in Sec.6.3.1, producing the aligned feature $\{A_h | h = 1, 2, \ldots, T\}$. Each of the aligned historical feature is separately combined with the current BEV feature B_t to form the input of the teacher ensemble $\{B'_h | h = 1, 2, \ldots, T\}$:

$$B'_h = [B_t, A_h], \tag{6.9}$$

where $[\cdot, \cdot]$ denotes feature concatenation. As shown in Figure 6.4, the concatenated features are individually sent to the teacher model to obtain their corresponding outputs. The final pseudo label computed as the average of all the outputs. As a consequence, the pseudo labels produced by the teacher temporal ensemble are more accurate than those generated from each individual frame, leading to more stable unsupervised consistency regularization. In addition, the temporal ensemble is only conducted during training, and thereby introducing no extra computational cost to inference.

Table 6.1. **Details of geometry transformations in the weak augmentation and intensity transformations in strong augmentation.**

Geometrical Augmentations	
Random Scaling	Randomly resizes the image.
Random Cropping	Randomly crops an region from the image.
Random Flipping	Horizontally flips the image with a probability of 0.5.
Intensity Augmentations	
Identity	Returns the original image.
AutoContrast	Maximizes (normalize) the image contrast.
RandEqualize	Equalize the image histogram.
RandSolarize	Inverts image pixels above a threshold from [1,256).
RandColor	Enhances the color balance of the image by [0.05, 0.95].
RandPosterize	Reduces the number of bits for each channel.
RandContrast	Adjusts the contrast of the image by [0.05, 0.95].
RandBrightness	Adjusts the brightness of the image by [0.05, 0.95].
RandSharpness	Adjusts the sharpness of the image by [0.05, 0.95].
CutOut	Masks out square regions of image.

6.3.4 Data Augmentation Strategies

Data augmentation strategies have been intensively studied for both semi-supervised 2D semantic segmentation (Yuan et al., 2021; Liu et al., 2022d) and fully supervised BEV perception tasks (Huang et al., 2021), yielding significant performance boosting. For instance, BEVDet (Huang et al., 2021) propose the data augmentation techniques for both BEV space and isolated image views to avoid over-fitting, which have been widely adopted by follow-up methods (Huang and Huang, 2022). Nonetheless, our experiments (Table 6.5, Table 6.6) show that these augmentations are less effective and even result in performance drop in semi-supervised BEV segmentation. We conjecture that these methods with over-distortions could hamper semi-supervised learning (Yuan et al., 2021), and the optimal augmentation strategies for semi-supervised BEV segmentation remain under-explored.

To mitigate the above issue, we perform a thorough investigation of a variety of data augmentation methods and design a new augmentation strategy to improve semi-supervised BEV segmentation. We classify all the data augmentation methods into two categories, *i.e.*, geometry augmentation (*e.g.*, random scaling, flipping, *etc.*) and intensity augmentation (*e.g.*, autocontrast, random color jittering, *etc.*), the details of which are illustrated in Table 6.1. We define the weak augmentation pipeline as the cascade of all the geometry augmentations and the strong augmentation pipeline as the integration of all the geometry augmentations and certain number of randomly selected intensity augmentations. During training, we follow the convention in semi-supervised learning and apply the weak and strong augmentation pipeline to the input of teacher and student model, respectively.

6.4 Experiments

6.4.1 Datasets and Metrics

NuScenes (Caesar et al., 2020) is a large scale autonomous driving benchmark, which is adopted for our main experiments. The sensor set for nuScenes contains 6 cameras, 1 LiDAR and 5 Radars, capturing 1000 driving scenes in Boston and Singapore, with 1.4M camera images in total. The whole dataset is divided into 850 scenes for training/validation and 150 for testing. In our experiments, we conduct semi-supervised learning using different proportions of labeled training images. Under complete training settings, we individually sample 10%, 20%, 30%, and 40% of all the training images as labeled data with the rest training images as unlabeled ones. Under the fast training setting, we randomly sample 5% and 20% of all the training images as labeled and unlabeled training images, respectively, to reduce training time.

As for evaluation, We adopt the Intersection-over-Union (IoU) as the metric. The ground truth includes seven categories that are common in autonomous driving: Drivable area, Lane, Vehicle, Pedestrian Crossing, Walkway, Stop Line, and Carpark. The lane category is formed by two map layers: Lane-Divider and Road-Divider. Following (Liu et al., 2022e), we maintain the overlaps between different categories and evaluate the binary segmentation result for each category separately.

6.4.2 Implementation Details

We use PETRv2 (Liu et al., 2022b) with segmentation head as our BEV segmentation network due to its simplicity and effectiveness. VoVNetV2 (Lee and Park, 2020) is employed as the backbone network and we conduct most of the experiments on it. Following BEVDet4D (Huang and Huang, 2022) and PETRv2 (Liu et al., 2022b), we randomly sample a previous frame from the range of $[3, 27]$ for student model during training, and sample the 15^{th} previous frame during inference. For teacher temporal ensemble, we randomly sample 3 previous frames from the range of $[3, 27]$ to include more historical information. We set the confidence thresholds to τ_{low}=0.2 and τ_{up}=0.8.

The model is trained for 24 epochs on 2 Nvidia A100 GPUs with a batch size of 4 using AdamW (Loshchilov and Hutter, 2017) optimizer with a weight decay of 0.01. The initial learning rate is 1×10^{-4} and is decayed following cosine annealing policy (Loshchilov and Hutter, 2016).

6.4.3 Ablation Studies

In this section, we conduct the ablations with VoVNet-99 backbone to explore the effectiveness of each proposed module. All the ablation experiments are trained under the fast training setting, *i.e.*, 5% labeled and 20% unlabeled training data for efficiency.

TABLE 6.2. **Threshold.** We compare results when differing the confidence thresholds.

Threshold	mIoU
$\tau_{low} = 0.4$ $\tau_{up} = 0.6$	49.1
$\tau_{low} = 0.3$ $\tau_{up} = 0.7$	50.0
$\tau_{low} = 0.2$ $\tau_{up} = 0.8$	**50.6**
$\tau_{low} = 0.1$ $\tau_{up} = 0.9$	50.5

TABLE 6.3. **Loss Functions**. Different loss functions used in unsupervised branch.

Loss Function	mIoU
$\mathcal{L}_{BCE}$	49.1
$\mathcal{L}_{Center}$	50.1
$\mathcal{L}_{Weight}$	18.3
$\mathcal{L}_{Conf}$	**50.6**

Impact of Loss Functions. We evaluate the performance of different confidence thresholds τ_{low} and τ_{up} in our confidence-aware unsupervised loss function $\mathcal{L}_u$ (Equation (6.7)). The results are shown in Table 6.2, where $\tau_{low} = 0.2, \tau_{up} = 0.8$ delivers the best result and is adopted by our final model. In addition, we also varify the effectiveness of our confidence-aware unsupervised loss function $\mathcal{L}_u$ by replacing it with other alternatives. Table 6.3 reports the comparison results, where '$\mathcal{L}_{BCE}$' is the vanilla binary cross entropy loss, '$\mathcal{L}_{Center}$' is the BEV Centerness loss (Xie et al., 2022), '$\mathcal{L}_{Conf}$' denotes the proposed confidence-aware cross entropy loss, and '$\mathcal{L}_{Weight}$' represents the weighted CE loss (Liu et al., 2022b). It shows that the proposed confidence-aware cross entropy loss outperforms all the other methods with significant margins.

TABLE 6.4. **Data Augmentation.** Ablative experiments on combinations of data augmentation strategies.

Teacher-Student	mIoU
weak-weak	48.9
strong-strong	50.3
weak-strong	**50.6**

Impact of Strong and Weak Data Augmentation. In our proposed BEV-S^4, weak and strong data augmentation pipelines are applied to the input of teacher and student model, respectively. To understand their impact, we experiment with different combination manner of these pipelines. As shown in Table 6.4, the approach 'X-Y' denotes to apply the 'X' and 'Y' augmentation pipelines to the input of teacher and student model, respectively. The comparison result shows that our adopted 'weak-strong' augmentation achieves the best performance, yielding 3.48% and 0.60% improvements compared to 'weak-weak' and 'strong-strong' baselines.

TABLE 6.5. **BEV Space Augmentation.** Effects of augmentations in BEV Space.

BEV Aug	mIoU
Flip	45.9
Scale	50.3
Rot	50.4

TABLE 6.6. **Augmentation View.** Effect of augmentations in different views and frames.

View	mIoU
all	**50.6**
frame-wise	50.2
image-wise	47.7

TABLE 6.7. **Augmentation Number.** Results of tuning number of intensity transformations.

k	mIoU
1	**50.6**
2	50.4
3	49.2

Analysis of Isolated View and BEV Space Augmentation. The isolated view and BEV space data augmentation methods proposed by (Huang et al., 2021) achieve considerable performance improvement for fully supervised BEV perception. To have a comprehensive understanding of these augmentation approaches, we evaluate their performance under the unsupervised setting. In Table 6.5, the method 'All' means to apply consistent data augmentation on all the 12 input images from current and previous frames. Following (Huang et al., 2021), 'View-wise' and 'Frame-wise' represent to apply different augmentation strategies for different views and different frames, respectively. In Table 6.6, 'Flip', 'Scale', and 'Rot' denote applying BEV space augmentation (Huang et al., 2021), *i.e.*, flipping, scaling and rotating together with our data augmentation strategies. The comparison results suggest that both isolated view and BEV space data augmentation fails to generalize to semi-supervised BEV segmentation task. We conjecture that these two augmentation strategy is too strong for semi-supervised learning. To partially verify this, we test the number of selected intensity transformations in strong augmentation. As shown in Table 6.7, more strong augmentations degrade the performance, confirming that over-distortions can indeed bring negative impact to the model. We thus set the number $k = 1$ in our experiments.

TABLE 6.8. **Teacher Temporal Ensemble.** Results of using different number of historical frames.

T	Intervals	mIoU
1	$[3, 27]$	49.2
2	$[3, 15], [15, 27]$	49.5
3	$[3, 11], [11, 19], [19, 27]$	**50.6**
4	$[3, 9], [9, 15], [15, 21], [21, 27]$	49.7

Effectiveness of Teacher Temporal Ensemble. To analyze the impact of teacher temporal ensemble, we evaluate the performance of using different number of historical frames as input. As shown in Table 6.8, T indicates the number of input historical frames and 'Intervals' denotes their sampling intervals. As a result, $T = 1$ represents the baseline without temporal ensemble. It can be seen that $T = 3$ achieves the best performance, which justifies the effectiveness of the teacher temporal ensemble.

6.4.4 Comparison

We evaluate the benefits of our BEV-S^4 method using various proportions of labeled data and compare its performance with the fully supervised baseline PETRv2 (Liu et al., 2022b). Table 6.9 presents the results on the nuScenes dataset in terms of per-category IoU and mIoU. Under different proportions of labeled data, our semi-supervised approach consistently outperforms the supervised baseline across all categories. The performance gap tends to widen as the amount of labeled data decreases, except for the 5% labeled set, where the unlabeled data is incomplete. Notably, with only 10% labeled data and 90% unlabeled data, our semi-supervised method achieves an mIoU improvement of 18.1%, increasing from 42.2% to 60.3%. This result is only 1.6% lower than the oracle, demonstrating that our method can serve as an effective solution when annotations are scarce.

Figure 6.5 illustrates the qualitative segmentation results. When compared to the supervised baseline, our semi-supervised model produces visually more accurate segmentation maps, further highlighting the efficacy of our approach.

6.5 Conclusion

In this paper, we present a semi-supervised pipeline designed to reduce the reliance on extensive labeled data for bird's-eye view (BEV) semantic segmentation. By leveraging both labeled and unlabeled data through semi-supervised learning, our proposed teacher-student dual training framework and data augmentation strategies enable the

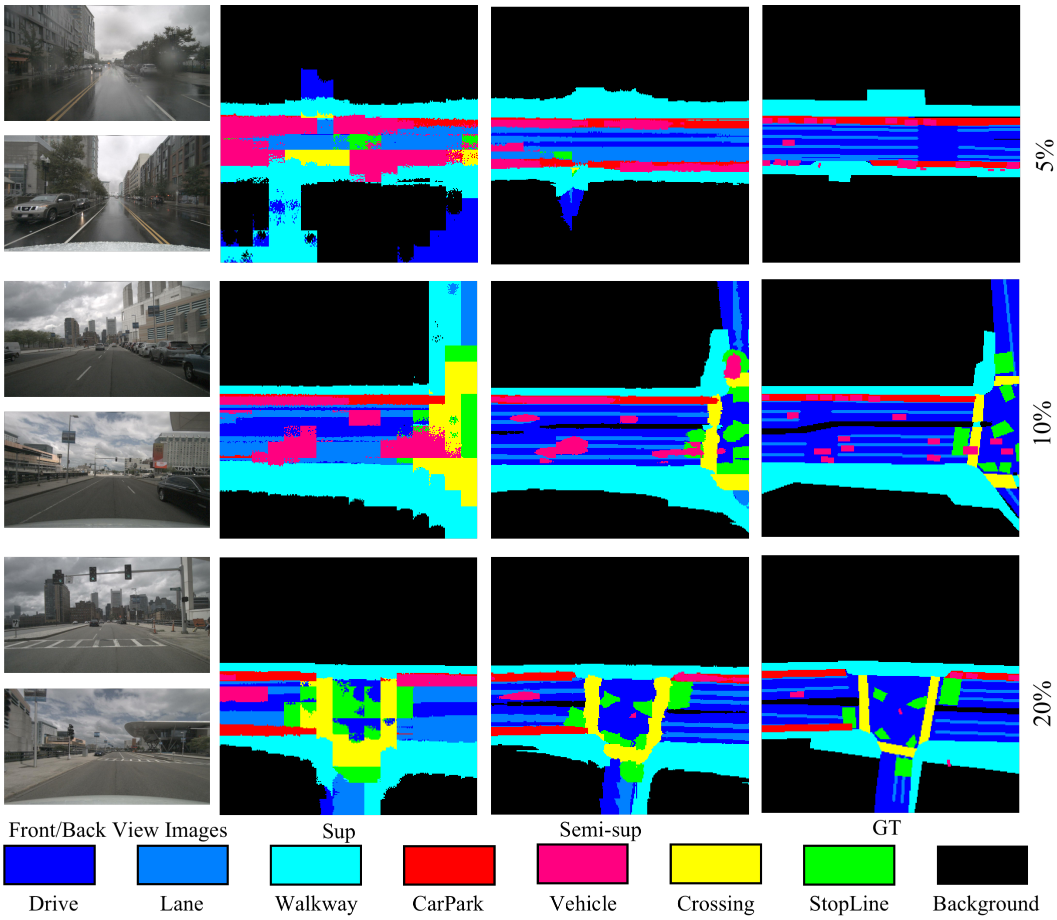


FIGURE 6.5. **Qualitative results on nuScenes.** We visually compare supervised baseline and our semi-supervised approach under various proportions of labeled data. Regions of different categories are distinguished with specified colors.

model to learn more robust representations via consistency training on pseudo-labels. Furthermore, we introduce a teacher temporal ensemble module, inspired by the inherent abundance of historical information in BEV tasks, to enhance performance further. Experimental results on the nuScenes dataset demonstrate that our BEV-S^4 method significantly improves performance by effectively utilizing unlabeled data. We anticipate that our approach will serve as a strong baseline for semi-supervised BEV semantic segmentation and offer valuable insights for future research on efficient BEV perception techniques.

Limitation and Discussion. In this work, our primary focus is on addressing the bird's-eye view (BEV) semantic segmentation using semi-supervised learning. However, we acknowledge that our approach has certain limitations. For instance, we have not implemented our method for other BEV tasks, such as object detection and 3D lane detection, which could potentially benefit from a joint framework. In the future, we plan to develop a unified semi-supervised pipeline for tackling multiple BEV tasks simultaneously. Another limitation of our BEV-S^4 method is the increased training time due to the incorporation of the teacher temporal ensemble, even though the inference speed remains unaffected. We aim to address this issue and further refine our approach in future research.

Table 6.9. **Compared with the fully supervised baseline on nuScenes dataset**. 'Proportion' represents the ratio of data used as labeled set during training. The fully supervised baseline 'Sup' is trained on labeled set and our semi-supervised approach 'Semi-sup' is trained on both labeled set and unlabeled set.

Proportions	Method	Drive	Lane	Vehicle	Crossing	Walkway	StopLine	Carpark	mIoU
5%	Sup	69.1	22.2	20.8	26.7	37.7	20.3	47.4	34.9
	Semi-sup	81.5	34.0	32.8	53.3	52.7	29.5	70.5	50.6
	Improv.	**+12.4**	**+11.8**	**+12.0**	**+26.6**	**+15.0**	**+9.2**	**+23.1**	**+15.7**
10%	Sup	74.8	26.1	25.7	36.5	43.8	23.0	65.2	42.2
	Semi-sup	86.4	45.6	44.1	65.5	62.3	43.9	74.1	60.3
	Improv.	**+11.6**	**+19.5**	**+18.4**	**+30.0**	**+18.5**	**+20.9**	**+8.9**	**+18.1**
20%	Sup	79.5	31.2	29.1	47.8	49.7	28.0	70.0	47.9
	Semi-sup	86.7	45.6	41.1	65.9	63.1	45.0	78.3	60.8
	Improv.	**+7.2**	**+14.4**	**+12.2**	**+18.1**	**+13.4**	**+17.0**	**+8.3**	**+12.9**
30%	Sup	82.2	35.4	32.5	54.4	55.1	32.4	74.1	52.3
	Semi-sup	86.7	45.4	40.7	66.8	63.0	44.6	79.6	61.0
	Improv.	**+4.5**	**+10.0**	**+8.2**	**+12.4**	**+7.9**	**+12.2**	**+5.5**	**+8.7**
40%	Sup	83.9	38.6	33.7	58.3	57.5	35.7	73.8	54.5
	Semi-sup	86.4	45.4	40.7	65.2	62.8	44.8	78.1	60.5
	Improv.	**+2.5**	**+6.8**	**+7.0**	**+6.9**	**+5.3**	**+9.1**	**+4.3**	**+6.0**
100%	Sup	87.1	46.6	40.9	68.0	64.3	46.6	79.6	61.9

Chapter 7

Conclusions

In recent years, image segmentation has emerged as a crucial perception task in the field of computer vision, particularly with the advent of large foundation models and AI-generated content transforming various aspects of our lives. As such, the development of practical segmentation methodologies has gained increased attention, with an emphasis on achieving accurate and robust results using fewer labeled samples and exploring techniques tailored for autonomous driving applications.

in this dissertation, we deliberately in solving several important tasks with careful designs, including few-shot semantic segmentation, unsupervised video object segmentation and semi-supervised bird's-eye-view semantic segmentation. The objective of our research is to develop novel methods and algorithms to improve segmentation performance in various scenarios.

Dynamic convolutions for few-shot segmentation. To adaptively leverage the relationships between the query set and support set, we proposed DRNet (Zhuge and Shen, 2021), which effectively exploited the category-level information from deep layers in the support branch through dynamic convolutions. Moreover, we introduced DRCNet (Gu et al., 2023), which explored contexts using a dynamic context module and a regional context module. These modules are responsible for extracting spatial information from query features and addressing ambiguous regions in query images, respectively, to produce more reliable results in few-shot semantic segmentation.

Learning motion and temporal cues for unsupervised video object segmentation. Temporal-based and motion-based unsupervised video object segmentation (UVOS) methods each offer their unique benefits. However, the potential of simultaneously exploiting both approaches remains largely untapped. In response, we introduced MTNet (Zhuge et al., 2023b), the first attempt at combining the advantages of both motion and temporal information for addressing UVOS. Our method achieved state-of-the-art results on multiple benchmarks while also providing real-time inference speeds on a 2080 Ti GPU, demonstrating its practicality for downstream tasks.

Semi-supervised bird's-eye-view semantic segmentation. Bird's-eye-view (BEV) perception is an increasingly important area of research, with BEV semantic segmentation gaining particular attention. However, the immense labor costs associated with ground truth labeling have hindered progress in the field. To address this challenge, we introduced Semi-BEV (Zhuge et al., 2023a), the first attempt at employing semi-supervised learning to tackle BEV semantic segmentation. Remarkably, we demonstrated that with only 10% of labeled data, our method can achieve results on par with a fully supervised approach trained on the entire dataset.

7.1 Future work

DRNet and DRCNet, as presented in Chapters 3 and 4, offer practical designs for tackling few-shot semantic segmentation using dynamic convolutions. While significant performance improvements have been observed, their effectiveness has only been demonstrated in 2D images. We hope to extend the application of few-shot segmentation to point cloud and autonomous driving scenarios, which would allow us to further address the reliance on labeled data, as we have made an initial attempt to do so in Chapter 6.

In the era of large foundation models and the explosion of unified multi-task methods, the future research direction in vision perception is filled with both challenges and opportunities. To begin with, current unified methods primarily focus on unifying tasks that share commonalities in their input, output modalities, and latent space. However, devising a universal solution for a wide range of differentiated tasks remains an open challenge. Furthermore, data scaling and model scaling play crucial roles in foundation models, but both aspects are computationally demanding. Thus, research on developing computation-friendly foundation models holds great potential and can significantly impact the future of vision perception.

Bibliography

Botach, Adam, Evgenii Zheltonozhskii, and Chaim Baskin (2022). “End-to-end referring video object segmentation with multimodal transformers”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4985–4995.

Boudiaf, Malik, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz (2021). “Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need?” In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 13979–13988.

Caesar, Holger, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom (2020). “nuscenes: A multimodal dataset for autonomous driving”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 11621–11631.

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (2020). “End-to-end object detection with transformers”. In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 213–229.

Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille (2017a). “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.4, pp. 834–848.

Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam (2017b). “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587*.

Chen, Xiaokang, Yuhui Yuan, Gang Zeng, and Jingdong Wang (2021). “Semi-supervised semantic segmentation with cross pseudo supervision”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 2613–2622.

Cheng, Bowen, Alex Schwing, and Alexander Kirillov (2021). “Per-pixel classification is not all you need for semantic segmentation”. In: *Proc. Advances in Neural Inf. Process. Syst.* 34, pp. 17864–17875.

Cheng, Ho Kei, Yu-Wing Tai, and Chi-Keung Tang (2021a). “Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 5559–5568.

Cheng, Ho Kei, Yu-Wing Tai, and Chi-Keung Tang (2021b). “Rethinking space-time networks with improved memory coverage for efficient video object segmentation”. In: *Proc. Advances in Neural Inf. Process. Syst.* 34, pp. 11781–11794.

Cho, Suhwan, Minhyeok Lee, Seunghoon Lee, and Sangyoun Lee (2022). "Domain Alignment and Temporal Aggregation for Unsupervised Video Object Segmentation". In: *arXiv preprint arXiv:2211.12036*.

Cubuk, Ekin D, Barret Zoph, Jonathon Shlens, and Quoc V Le (2020). "Randaugment: Practical automated data augmentation with a reduced search space". In: pp. 702–703.

Dai, Kenan, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang (2020). "High-performance long-term tracking with meta-updater". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 6298–6307.

Dong, Nanqing and Eric P Xing (2018). "Few-Shot Semantic Segmentation with Prototype Learning." In: *Proc. British Machine Vis. Conf.* Vol. 3. 4.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *Proc. Int. Conf. Learn. Repre.*

Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman (2010). "The pascal visual object classes (voc) challenge". In: *Int. J. Comput. Vision* 88.2, pp. 303–338.

Fan, Deng-Ping, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen (2019). "Shifting more attention to video salient object detection". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 8554–8564.

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

Gavrilyuk, Kirill, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek (2018). "Actor and action video segmentation from a sentence". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 5958–5966.

Ge, Chongjian, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo (2023). "MetaBEV: Solving Sensor Failures for BEV Detection and Map Segmentation". In: *arXiv preprint arXiv:2304.09801*.

Gu, Hongyu, Yunzhi Zhuge, Lu Zhang, Jinqing Qi, and Huchuan Lu (2023). "Few-shot Semantic Segmentation by Exploiting Dynamic and Regional Contexts". In: *Proc. IEEE Int. Conf. Multi. Expo.*

Gu, Jiaqi, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z Pan (2022). "Multi-scale high-resolution vision transformer for semantic segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 12094–12103.

Gu, Yuchao, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu (2020). "Pyramid constrained self-attention network for fast video salient object detection". In: *Proc. AAAI Conf. Artificial Intell.* Vol. 34. 07, pp. 10869–10876.

Guo, Jianyuan, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu (2022). "Cmt: Convolutional neural networks meet vision transformers". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 12175–12185.

Guo, Yanming, Yu Liu, Theodoros Georgiou, and Michael S Lew (2018). "A review of semantic segmentation using deep neural networks". In: *Int. J. Mul. Inf. Retri.* 7, pp. 87–93.

Gupta, Surbhi, R Sangeeta, Ravi Shankar Mishra, Gaurav Singal, Tapas Badal, and Deepak Garg (2020). "Corridor segmentation for automatic robot navigation in indoor environment using edge devices". In: *Computer Networks* 178, p. 107374.

Hao, Shijie, Yuan Zhou, and Yanrong Guo (2020). "A brief survey on semantic segmentation with deep learning". In: *J. Neuro.* 406, pp. 302–321.

Hariharan, Bharath, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik (2011). "Semantic contours from inverse detectors". In: *Proc. IEEE Int. Conf. Comp. Vis.* IEEE, pp. 991–998.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 770–778.

He, Ruifei, Jihan Yang, and Xiaojuan Qi (2021). "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 6930–6940.

Hendrycks, Dan and Thomas Dietterich (2019). "Benchmarking neural network robustness to common corruptions and perturbations". In: *arXiv preprint arXiv:1903.12261*.

Heo, Yuk, Yeong Jun Koh, and Chang-Su Kim (2021). "Guided interactive video object segmentation using reliability-based attention maps". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 7322–7330.

Hong, Sunghwan, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim (2022). "Cost aggregation with 4d convolutional swin transformer for few-shot segmentation". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 108–126.

Hou, Qibin, Li Zhang, Ming-Ming Cheng, and Jiashi Feng (2020). "Strip pooling: Rethinking spatial pooling for scene parsing". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4003–4012.

Hu, Yihan, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. (2023). "Goal-oriented Autonomous Driving". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Huang, Junjie and Guan Huang (2022). "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection". In: *arXiv preprint arXiv:2203.17054*.

Huang, Junjie, Guan Huang, Zheng Zhu, and Dalong Du (2021). "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view". In: *arXiv preprint arXiv:2112.11790*.

Huang, Zilong, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu (2019). "Ccnet: Criss-cross attention for semantic segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 603–612.

Jampani, Varun, Raghudeep Gadde, and Peter V Gehler (2017). "Video propagation networks". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 451–461.

Jang, Won-Dong and Chang-Su Kim (2017). "Online video object segmentation via convolutional trident network". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 5849–5858.

Ji, Ge-Peng, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao (2021). "Full-duplex strategy for video object segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 4922–4933.

Kang, Bingyi, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell (2019). "Few-shot object detection via feature reweighting". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 8420–8429.

Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele (2019). "Lucid data dreaming for video object segmentation". In: *Int. J. Comput. Vision* 127.9, pp. 1175–1197.

Khoreva, Anna, Anna Rohrbach, and Bernt Schiele (2019). "Video object segmentation with language referring expressions". In: *Proc. Asian Conf. Comp. Vis.* Springer, pp. 123–141.

Kim, Jongmok, Jooyoung Jang, Hyunwoo Park, and SeongAh Jeong (2020). "Structured consistency loss for semi-supervised semantic segmentation". In: *arXiv preprint arXiv:2001.04647*.

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

Krähenbühl, Philipp and Vladlen Koltun (2011). "Efficient inference in fully connected crfs with gaussian edge potentials". In: *Proc. Advances in Neural Inf. Process. Syst.* 24.

Lang, Chunbo, Gong Cheng, Binfei Tu, and Junwei Han (2022). "Learning what not to segment: A new perspective on few-shot segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 8057–8067.

Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce (2006). "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Vol. 2. IEEE, pp. 2169–2178.

Lee, Youngwan and Jongyoul Park (2020). "Centermask: Real-time anchor-free instance segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 13906–13915.

Li, Fuxin, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg (2013). "Video segmentation by tracking many figure-ground segments". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 2192–2199.

Li, Gen, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim (2021). "Adaptive Prototype Learning and Allocation for Few-Shot Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 8334–8343.

Li, Xiang, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang (2020). "Fss-1000: A 1000-class dataset for few-shot segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 2869–2878.

Li, Zhiqi, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai (2022). "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers". In: *Proc. Eur. Conf. Comp. Vis.*

Liang, Chen, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang (2023). "Local-global context aware transformer for language-guided video segmentation". In: *IEEE Trans. Pattern Anal. Mach. Intell.*

Liang, Yongqing, Xin Li, Navid Jafari, and Jim Chen (2020). "Video object segmentation with adaptive feature bank and uncertain-region refinement". In: *Proc. Advances in Neural Inf. Process. Syst.* 33, pp. 3430–3441.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick (2014). "Microsoft COCO: Common objects in context". In: *Proc. Eur. Conf. Comp. Vis.* Pp. 740–755.

Liu, Daizong, Dongdong Yu, Changhu Wang, and Pan Zhou (2021a). "F2net: Learning to focus on the foreground for unsupervised video object segmentation". In: *Proc. AAAI Conf. Artificial Intell.* Vol. 35. 3, pp. 2109–2117.

Liu, Jianbo, Junjun He, Yu Qiao, Jimmy S Ren, and Hongsheng Li (2020a). "Learning to predict context-adaptive convolution for semantic segmentation". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 769–786.

Liu, Jiang-Jiang, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang (2019). "A simple pooling-based design for real-time salient object detection". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 3917–3926.

Liu, Jie, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves (2022a). "Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 11553–11562.

Liu, Weide, Chi Zhang, Guosheng Lin, and Fayao Liu (2020b). "CRNet: Cross-Reference Networks for Few-Shot Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4165–4173.

Liu, Yingfei, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun (2022b). "Petrv2: A unified framework for 3d perception from multi-camera images". In: *arXiv preprint arXiv:2206.01256.*

Liu, Yongfei, Xiangyi Zhang, Songyang Zhang, and Xuming He (2020c). "Part-aware prototype network for few-shot semantic segmentation". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 142–158.

Liu, Yuanwei, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao (2022c). "Learning Non-Target Knowledge for Few-Shot Semantic Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 11573–11582.

Liu, Yuyuan, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro (2022d). "Perturbed and strict mean teachers for semi-supervised semantic segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4258–4267.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo (2021b). "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 10012–10022.

Liu, Zhijian, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han (2022e). "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation". In: *arXiv preprint arXiv:2205.13542.*

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie (2022f). "A convnet for the 2020s". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 11976–11986.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell (2015). "Fully convolutional networks for semantic segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 3431–3440.

Loshchilov, Ilya and Frank Hutter (2016). "Sgdr: Stochastic gradient descent with warm restarts". In: *arXiv preprint arXiv:1608.03983.*

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

Lu, Xiankai, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli (2019). "See more, know more: Unsupervised video object segmentation with co-attention siamese networks". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 3623–3632.

Lu, Zhihe, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang (2021). "Simpler Is Better: Few-Shot Semantic Segmentation With Classifier Weight Transformer". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 8741–8750.

Mao B., Zhang X. Wang L. Zhang Q. Xiang S. Pan C. (2022). "Learning from the Target: Dual Prototype Network for Few Shot Semantic Segmentation". In: *Proc. AAAI Conf. Artificial Intell.* Pp. 1953–1961.

Mehta, Sachin, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi (2018). "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation". In: *Proc. Eur. Conf. Comp. Vis.* Pp. 552–568.

Micikevicius, Paulius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. (2017). "Mixed precision training". In: *arXiv preprint arXiv:1710.03740.*

Min, Juhong, Dahyun Kang, and Minsu Cho (2021). "Hypercorrelation squeeze for few-shot segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 6941–6952.

Ng, Mong H, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez (2020). "BEV-Seg: Bird's Eye View Semantic Segmentation Using Geometry and Semantic Point Cloud". In: *arXiv preprint arXiv:2006.11436.*

Nguyen, Khoi and Sinisa Todorovic (2019). "Feature weighting and boosting for few-shot segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 622–631.

Ochs, Peter, Jitendra Malik, and Thomas Brox (2013). “Segmentation of moving objects by long term video analysis”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.6, pp. 1187–1200.

Oh, Seoung Wug, Joon-Young Lee, Ning Xu, and Seon Joo Kim (2019a). “Fast user-guided video object segmentation by interaction-and-propagation networks”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 5247–5256.

Oh, Seoung Wug, Joon-Young Lee, Ning Xu, and Seon Joo Kim (2019b). “Video object segmentation using space-time memory networks”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 9226–9235.

Olsson, Viktor, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson (2021). “Classmix: Segmentation-based data augmentation for semi-supervised learning”. In: *Proc. IEEE Win. Conf. App. Comp. Vis.* Pp. 1369–1378.

Ouali, Yassine, Céline Hudelot, and Myriam Tami (2020). “Semi-supervised semantic segmentation with cross-consistency training”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 12674–12684.

Pan, Bowen, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou (2020). “Cross-view semantic segmentation for sensing surroundings”. In: *IEEE Robot. Auto. Letters* 5.3, pp. 4867–4873.

Papazoglou, Anestis and Vittorio Ferrari (2013). “Fast object segmentation in unconstrained video”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 1777–1784.

Park, Jinhyung, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan (2022a). “Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection”. In: *Proc. Int. Conf. Learn. Repre.*

Park, Jinhyung, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan (2022b). “Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection”. In: *arXiv preprint arXiv:2210.02443*.

Pei, Gensheng, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang (2022). “Hierarchical feature alignment network for unsupervised video object segmentation”. In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 596–613.

Peng, Chao, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun (2017). “Large kernel matters–improve semantic segmentation by global convolutional network”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4353–4361.

Perazzi, Federico, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung (2017). “Learning video object segmentation from static images”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 2663–2672.

Perazzi, Federico, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung (2016). “A benchmark dataset and evaluation methodology for video object segmentation”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 724–732.

Philion, Jonah and Sanja Fidler (2020). “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d”. In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 194–210.

Prest, Alessandro, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari (2012). “Learning object class detectors from weakly annotated video”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, pp. 3282–3289.

Rakelly, Kate, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine (2018). “Conditional networks for few-shot semantic segmentation”. In.

Ren, Sucheng, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He (2021). “Reciprocal transformations for unsupervised video object segmentation”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 15455–15464.

Roddick, Thomas and Roberto Cipolla (2020). “Predicting semantic map representations from images using pyramid occupancy networks”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 11138–11147.

Seo, Seonguk, Joon-Young Lee, and Bohyung Han (2020). “Urvos: Unified referring video object segmentation network with a large-scale benchmark”. In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 208–223.

Seong, Hongje, Junhyuk Hyun, and Euntai Kim (2020). “Kernelized memory network for video object segmentation”. In: *Proc. Eur. Conf. Comp. Vis.* Springer.

Shaban, Amirreza, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots (2017). “One-shot learning for semantic segmentation”. In: *arXiv preprint arXiv:1709.03410*.

Siam, Mennatullah, Boris N Oreshkin, and Martin Jagersand (2019). “AMP: Adaptive masked proxies for few-shot segmentation”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 5249–5258.

Simonyan, Karen and Andrew Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556*.

Singh, Amitojdeep, Sourya Sengupta, and Vasudevan Lakshminarayanan (2020). “Explainable deep learning models in medical image analysis”. In: *Journal of Imaging* 6.6, p. 52.

Snell, Jake, Kevin Swersky, and Richard Zemel (2017). “Prototypical networks for few-shot learning”. In: *Proc. Advances in Neural Inf. Process. Syst.* 30.

Song, Hongmei, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam (2018). “Pyramid dilated deeper convlstm for video salient object detection”. In: *Proc. Eur. Conf. Comp. Vis.* Pp. 715–731.

Teed, Zachary and Jia Deng (2020). “Raft: Recurrent all-pairs field transforms for optical flow”. In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 402–419.

Tian, Zhi, Chunhua Shen, and Hao Chen (2020). “Conditional Convolutions for Instance Segmentation”. In.

Tian, Zhi, Bowen Zhang, Hao Chen, and Chunhua Shen (2022). “Instance and panoptic segmentation using conditional convolutions”. In: *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 669–680.

Tian, Zhuotao, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia (2020). “Prior guided feature enrichment network for few-shot segmentation”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 44.2, pp. 1050–1065.

Tokmakov, Pavel, Karteek Alahari, and Cordelia Schmid (2017). “Learning video object segmentation with visual memory”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 4481–4490.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Proc. Advances in Neural Inf. Process. Syst.* 30.

Wang, Haochen, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen (2020). “Few-Shot Semantic Segmentation with Democratic Attention Networks”. In: *Proc. Eur. Conf. Comp. Vis.*

Wang, Kaixin, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng (2019a). “Panet: Few-shot image semantic segmentation with prototype alignment”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 9197–9206.

Wang, Tiantian, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji (2018a). “Detect globally, refine locally: A novel approach to saliency detection”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 3127–3135.

Wang, Wenguan, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao (2019b). “Zero-shot video object segmentation via attentive graph neural networks”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 9236–9245.

Wang, Wenguan, Jianbing Shen, and Ling Shao (2015). “Consistent video saliency using local gradient flow optimization and global refinement”. In: *IEEE Trans. Image Process.* 24.11, pp. 4185–4196.

Wang, Wenguan, Jianbing Shen, Ruigang Yang, and Fatih Porikli (2017). “Saliency-aware video object segmentation”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.1, pp. 20–33.

Wang, Wenguan, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling (2019c). “Learning unsupervised video object segmentation through visual attention”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 3064–3074.

Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He (2018b). “Non-local neural networks”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 7794–7803.

Wang, Yuchao, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le (2022). “Semi-supervised semantic segmentation using unreliable pseudo-labels”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4248–4257.

Wikipedia contributors (2023). *Tesla Autopilot.* Accessed: 2023-04-30. URL: https://en.wikipedia.org/wiki/Tesla_Autopilot.

Wu, Jiannan, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo (2022). “Language as queries for referring video object segmentation”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4974–4984.

Wu, Zhonghua, Xiangxi Shi, Guosheng Lin, and Jianfei Cai (2021). “Learning meta-class memory for few-shot semantic segmentation”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 517–526.

Xiao, Huaxin, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang (2018). "Monet: Deep motion exploitation for video object segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 1140–1148.

Xie, Enze, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo (2021a). "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Proc. Advances in Neural Inf. Process. Syst.* 34, pp. 12077–12090.

Xie, Enze, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez (2022). "M^2-BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Birds-Eye View Representation". In: *arXiv preprint arXiv:2204.05088.*

Xie, Guo-Sen, Jie Liu, Huan Xiong, and Ling Shao (2021b). "Scale-Aware Graph Neural Network for Few-Shot Semantic Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 5475–5484.

Xie, Guo-Sen, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao (2021c). "Few-Shot Semantic Segmentation With Cyclic Memory Network". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 7293–7302.

Xie, Haozhe, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun (2021d). "Efficient regional memory network for video object segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 1286–1295.

Xie, Shaoyuan, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu (2023). "RoboBEV: Towards Robust Bird's Eye View Perception under Corruptions". In: *arXiv preprint arXiv:2304.06719.*

Xu, Ning, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang (2018). "Youtube-vos: Sequence-to-sequence video object segmentation". In: *Proc. Eur. Conf. Comp. Vis.* Pp. 585–601.

Xu, Rui, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy (2019). "Deep flow-guided video inpainting". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 3723–3732.

Xue, Tianfan, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman (2019). "Video enhancement with task-oriented flow". In: *Int. J. Comput. Vision* 127, pp. 1106–1125.

Yan, Pengxiang, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin (2019). "Semi-supervised video salient object detection using pseudo-labels". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7284–7293.

Yang, Boyu, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye (2020a). "Prototype mixture models for few-shot semantic segmentation". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 763–778.

Yang, Lihe, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao (2021a). "Mining latent classes for few-shot segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 8721–8730.

Yang, Lihe, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao (2022). "St++: Make self-training work better for semi-supervised semantic segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 4268–4277.

Yang, Shu, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, and Xiaoxing Zhang (2021b). "Learning motion-appearance co-attention for zero-shot video object segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 1564–1573.

Yang, Xianghui, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou (2020b). "BriNet: Towards Bridging the Intra-class and Inter-class Gaps in One-Shot Segmentation". In.

Yang, Zhao, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr (2019). "Anchor diffusion for unsupervised video object segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 931–940.

Yang, Zongxin, Yunchao Wei, and Yi Yang (2020). "Collaborative video object segmentation by foreground-background integration". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 332–348.

Yin, Zhaoyuan, Jia Zheng, Weixin Luo, Shenhan Qian, Hanling Zhang, and Shenghua Gao (2021). "Learning to recommend frame for interactive video object segmentation in the wild". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 15445–15454.

Yu, Changqian, Yifan Liu, Changxin Gao, Chunhua Shen, and Nong Sang (2020). "Representative graph neural network". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 379–396.

Yu, Changqian, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang (2018). "Learning a discriminative feature network for semantic segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 1857–1866.

Yu, Fisher and Vladlen Koltun (2015). "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122.*

Yuan, Jianlong, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li (2021). "A Simple Baseline for Semi-supervised Semantic Segmentation with Strong Data Augmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 8229–8238.

Yun, Sangdoo, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo (2019). "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 6023–6032.

Zeng, Yanhong, Jianlong Fu, and Hongyang Chao (2020). "Learning joint spatial-temporal transformations for video inpainting". In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 528–543.

Zhang, Bingfeng, Jimin Xiao, and Terry Qin (2021). "Self-Guided and Cross-Guided Learning for Few-Shot Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 8312–8321.

Zhang, Chi, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao (2019a). "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 9587–9595.

Zhang, Chi, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen (2019b). “Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 5217–5226.

Zhang, Gengwei, Guoliang Kang, Yi Yang, and Yunchao Wei (2021a). “Few-shot segmentation via cycle-consistent transformer”. In: *Proc. Advances in Neural Inf. Process. Syst.* 34, pp. 21984–21996.

Zhang, Hongyi, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz (2017). “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412.*

Zhang, Kaihua, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu (2021b). “Deep transport network for unsupervised video object segmentation”. In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 8781–8790.

Zhang, Miao, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo (2021c). “Dynamic context-sensitive filtering network for video salient object detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1553–1563.

Zhang, Pingping, Luyao Wang, Dong Wang, Huchuan Lu, and Chunhua Shen (2018). “Agile Amulet: Real-Time Salient Object Detection with Contextual Attention”. In: *arXiv:1802.06960.*

Zhang, Shan, Tianyi Wu, Sitong Wu, and Guodong Guo (2022). “Catrans: context and affinity transformer for few-shot segmentation”. In: *Proc. Int. Joint Conf. Artificial Intell.*

Zhang, Xiaolin, Yunchao Wei, Yi Yang, and Thomas S Huang (2020a). “Sg-one: Similarity guidance network for one-shot semantic segmentation”. In: *IEEE Trans. Cybern.* 50.9, pp. 3855–3865.

Zhang, Yizhuo, Zhirong Wu, Houwen Peng, and Stephen Lin (2020b). “A transductive approach for video object segmentation”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 6949–6958.

Zhang, Yunhua, Lijun Wang, Dong Wang, Jinqing Qi, and Huchuan Lu (2021d). “Learning regression and verification networks for robust long-term tracking”. In: *Int. J. Comput. Vision* 129.9, pp. 2536–2547.

Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia (2017). “Pyramid scene parsing network”. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 2881–2890.

Zhao, Zhen, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang (2022). “Augmentation Matters: A Simple-yet-Effective Approach to Semi-supervised Semantic Segmentation”. In: *arXiv preprint arXiv:2212.04976.*

Zhen, Mingmin, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan (2020). “Learning discriminative feature with crf for unsupervised video object segmentation”. In: *Proc. Eur. Conf. Comp. Vis.* Springer, pp. 445–462.

Zheng, Sixiao, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. (2021). “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 6881–6890.

Zhou, Brady and Philipp Krähenbühl (2022). "Cross-view Transformers for real-time Map-view Semantic Segmentation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 13760–13769.

Zhou, Peng, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis (2020a). "Generate, segment, and refine: Towards generic manipulation segmentation". In: *Proc. AAAI Conf. Artificial Intell.* Vol. 34. 07, pp. 13058–13065.

Zhou, Tianfei, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao (2020b). "Motion-attentive transition for zero-shot video object segmentation". In: *Proc. AAAI Conf. Artificial Intell.* Vol. 34. 07, pp. 13066–13073.

Zhou, Yimin, Ling Tian, Ce Zhu, Xin Jin, and Yu Sun (2019). "Video coding optimization for virtual reality 360-degree source". In: *IEEE Journal of Selected Topics in Signal Processing* 14.1, pp. 118–129.

Zhu, Xizhou, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei (2017). "Deep feature flow for video recognition". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pp. 2349–2358.

Zhu, Zhen, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai (2019). "Asymmetric non-local neural networks for semantic segmentation". In: *Proc. IEEE Int. Conf. Comp. Vis.* Pp. 593–602.

Zhuge, Yunzhi, Hongyu Gu, Lu Zhang, Jiqing Qi, and Huchuan Lu (2023a). "BEVS^4: Semi-supervised Bird's-eye-view Semantic Segmentation". In: *Under review of ACM MM 2023.*

Zhuge, Yunzhi, Hongyu Gu, Lu Zhang, Jiqing Qi, and Huchuan Lu (2023b). "Learning Motion and Temporal Cues for Unsupervised Video Object Segmentation". In: *Under review of ACM MM 2023.*

Zhuge, Yunzhi and Chunhua Shen (2021). "Deep reasoning network for few-shot semantic segmentation". In: *Proc. ACM Int. Conf. Multimedia*, pp. 5344–5352.

Zou, Yuliang, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister (2020). "Pseudoseg: Designing pseudo labels for semantic segmentation". In: *Proc. Int. Conf. Learn. Repre.*