國立陽明交通大學

電機資訊國際學位學程

碩士論文

EECS International Graduate Program

National Yang Ming Chiao Tung University

Master Thesis

域自適應邊緣引導的 DETR 用於目標檢測

Domain Adaptive Edge-guided DETR for Object Detection

研究生：馬大衛 (Marco Galindo)

指導教授：帥宏翰 (Hong-Han Shuai)

聯合指導：鄭文皇 (Wen-Huang Cheng)

中華民國一一三年一月

January 2024

域自適應邊緣引導的 DETR 用於目標檢測

Domain Adaptive Edge-guided DETR for Object Detection

研究生：馬大衛　　Student : Marco Galindo

指導教授：帥宏翰　教授　　Advisor : Hong-Han Shuai

聯合指導：鄭文皇　教授　　Co-Advisor : Wen-Huang Cheng

國立陽明交通大學

電機資訊國際學位學程

碩士論文

A Thesis
Submitted to EECS International Graduate Program
National Yang Ming Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Electrical Engineering and Computer Science

January 2024
Taiwan, Republic of China

中華民國一一三年一月

Authorization Form

Validation Form

Acknowledgments

Before anything, I am profoundly grateful to God for providing me with the strength, wisdom, and perseverance needed to complete this degree. My faith has been a constant source of guidance and comfort throughout this academic journey. I want to express my heartfelt gratitude to my family for their unwavering support, love, and understanding, even if they are far from me now. Their encouragement provided the emotional strength needed to persevere through the challenges of this academic journey. A special thanks to my friends and peers who have been a source of inspiration, motivation, and camaraderie. Their intellectual exchange and camaraderie have enriched my academic experience. Especially, I would like to thank 劉厚毅 for mentoring me during this journey and providing me with the sufficient knowledge to complete this thesis. Finally, I am deeply thankful to my supervisors, 鄭文皇 and 帥宏翰 , for their guidance, support, and providing me with key resources during my research process. Their expertise and unwavering commitment to fostering a stimulating academic environment have been pivotal in shaping the trajectory of my academic journey.

This thesis is a culmination of collective efforts, and I am grateful to everyone who has been a part of this academic endeavor. Your support has been invaluable, and I am truly thankful for the contributions each of you has made.

域自適應邊緣引導的 DETR 用於目標檢測

研究生：馬大衛　　指導教授：帥宏翰, 鄭文皇

電機資訊國際學位學程

國立陽明交通大學

摘要

這項研究介紹了一種針對提升DETR框架泛化能力而設計的新型即插即用模塊。利用該模塊，我們在競爭激烈的最先進模型中實現了實質的1.9%性能提升，將檢測準確性從46.8%提高到48.7%。我們的方法巧妙地利用了領域不可知的信息來優化解碼器查詢，從而有助於提升檢測性能。本研究揭示了整合專門模塊以提升DETR等物體檢測模型能力的潛力，對於在不同領域中提升性能具有深遠的影響。

關鍵詞: DETR, Domain Adaptation, Object Detection, Cityscapes

Domain Adaptive Edge-guided DETR for Object Detection

Student : Marco Galindo
Advisor : Hong-Han Shuai
Co-Advisor : Wen-Huang Cheng

EECS International Graduate Program
National Yang Ming Chiao Tung University

Abstract

This research introduces a novel plug-and-play module tailored for enhancing the generalization capacity of the DETR framework. Leveraging this module, we achieved a substantial 1.9% performance improvement in a competitive state-of-the-art model, elevating detection accuracy from 46.8% to 48.7%. Our approach strategically exploits domain-agnostic information to refine decoder queries, thereby contributing to improved detection performance. This study sheds light on the potential of integrating specialized modules to enhance the capabilities of object detection models like DETR, with implications for advancing performance across diverse domains.

Keywords: DETR, Domain Adaptation, Object Detection, Cityscapes

Table of Contents

List of Figures

List of Tables

Chapter I
Introduction

In recent years, computer vision algorithms have made remarkable progress, especially in object detection tasks within complex real-world scenarios like autonomous driving. Yet, despite their impressive performance within their training domains, they often stumble when faced with even slight variations in the environment. This lack of generalizability hampers their practicality. For instance, a model excelling on city roads under clear skies might falter during rainy or foggy conditions when this labeled data is scarce or absent. This is where unsupervised domain adaptation enters the stage. In this work, we explore the crucial challenge of training models on one fully labeled domain and deploying them seamlessly in unlabeled domains. Our focus is on enhancing the proficiency of computer vision algorithms in diverse, unpredictable environments. We delve into the intricacies of adapting these algorithms to new, uncharted territory, ultimately aiming to bridge the gap between training data and real-world application.

1.1 Settings for Object Adaptation

In practice, models that are trained on one domain, tend to not perform as well when this domain is even slightly changed. Take the example of

pedestrian detection for a self-driving car. If said car's model was trained on a dataset containing only nighttime images, the performance will suffer greatly when trying to detect the same pedestrians on daytime, rainy or foggy days. The ability for a model to perform well in different domains while only being trained on one is the ultimate goal of domain adaptation and why it is so important, given it saves computation and money.

In practical scenarios, machine learning models trained on a specific domain tend to experience a notable decline in performance when exposed to even minor variations within that domain. To illustrate, consider the scenario of pedestrian detection in a self-driving car. If the model is exclusively trained on a dataset consisting of nighttime images, its efficacy markedly diminishes when faced with the challenge of detecting pedestrians during daytime, in rainy conditions, or in the presence of fog. The overarching objective of domain adaptation is to find a model with the capability to excel across a spectrum of domains while being trained exclusively on one. This pursuit is of paramount importance as it not only enhances performance across diverse scenarios but also results in substantial savings in computation and financial resources.

1.2 Motivation

In the domain adaptation context, it is established that deeper layers of a model harbor more domain-specific information, while shallower layers contain more domain-agnostic information. In line with this understanding, our objective is to construct a module that enhances low-level informa-

tion within the deeper segments of the DETR structure. Furthermore, our approach extends beyond the utilization of information solely from shallow sections; it incorporates additional low-level details from a lightweight external network designed to extract edges from input images.

The inclusion of edge information in our research is motivated by the recognition that edges can be effectively extracted from the shallow layers of a model, rendering them a valuable source of domain-invariant information. Importantly, edges play a pivotal role in addressing domain shifts, as they often exhibit consistency across diverse domains. For example, the edges of an object, such as a car, tend to remain relatively stable whether captured on a sunny or rainy day. We hypothesize that feeding the DETR model with this more object-related details while mitigating background variations could enhance its performance in the face of domain shifts.

1.3 Our Contributions

In contrast to numerous state-of-the-art domain adaptation methods that predominantly employ techniques such as adversarial learning or knowledge distillation in their frameworks, we propose an alternative strategy. Our approach involves the incorporation of external information from an edge extractor to harness domain-agnostic knowledge that may not be explicitly present in the model. Importantly, we embed this strategy within a modular framework that seamlessly integrates with the DETR architecture. Our method presents several notable contributions in comparison to existing approaches, which are outlined as follows:

1. We introduce a plug-and-play model specifically engineered to augment the generalization capacity of DETR towards novel domains.

2. By harnessing information from distinct segments of the DETR structure, we amplify domain-agnostic features embedded within shallower sections.

3. Our proposed approach demonstrates competitive performance, as evidenced by its achievement relative to top-performing models on the Cityscapes to Foggyscapes datasets.

Chapter II
Related Works

In this section, we detail some previous work done on similar areas of study.

2.1 Object Detection

Object detection, one of the main tasks within the field of computer vision, involves the identification and localization of objects within an image or video frame. The significance of object detection extends across diverse applications, including autonomous vehicles, surveillance systems, image retrieval, and augmented reality. As computer vision has evolved, so has the complexity of object detection methodologies, ranging from early rule-based systems to the latest deep learning approaches. In the early stages of object detection research, CNN-based detectors emerged as the prevailing solution. These detectors were categorized into either two-stage or one-stage configurations. Pioneering this field, the R-CNN introduced the two-stage detection approach. Initially, it conducted a region proposal search, identifying potential regions of interest, and subsequently refined these proposals, ultimately discerning the objects contained within them. The R-CNN laid the foundation for subsequent advancements in two-stage object detection methodologies. The Faster R-CNN further enhances the

framework by proposing a Region Proposal Network (RPN) to replace the selective search stage.

On the contrary, one-stage networks are designed to simultaneously predict bounding boxes and object classes. Although these methods generally offer faster processing compared to two-stage detectors, they may exhibit a slight decrease in accuracy. The YOLO series [1] represents a widely acclaimed set of algorithms that significantly advanced this approach. Subsequently, with a pivotal breakthrough in deep learning arising from the introduction of transformer architectures [2], an efficient one-stage detector known as DETR [3] was conceived. This model conceptualizes the object detection task as a set of predictions, utilizing a Transformer network in conjunction with a CNN backbone to encode relationships among set elements. A key innovation lies in the introduction of the bipartite loss, a mechanism employed to align network predictions with ground-truth boxes. While this method demonstrated remarkable performance and removed the need for problem-specific modifications, it encountered challenges such as prolonged convergence times and difficulties in detecting small objects. Shortly thereafter, the Deformable DETR [4] emerged as a noteworthy advancement over the baseline model. The integration of a sparse attention module played a pivotal role in significantly reducing the convergence time by substantial margins. Owing to its outstanding performance and acceptable model size, this model is becoming a popular starting point for domain adaptation for object detection algorithms.

In the context of domain adaptation, it is acknowledged that deeper

layers encapsulate domain-specific knowledge, while shallower layers contribute domain-agnostic information [5]. Taking this into consideration, some approaches, such as those presented by Yu et al. [6], incorporate alignment solutions across multiple sections of the DETR architecture. In our proposed methodology, we strive to enhance the model's generalization not merely by aligning specific sections but by combining information from diverse segments. Specifically, we integrate general features from both the backbone and an auxiliary edge-extractor network, along with higher-level features derived from the encoder. This mixture of information is then employed to guide the queries within the decoder section.

2.2 Domain Adaptive Object Detection

In the domain adaptation field, we consider two distinct domains: the source domain and the target domain. These domains are closely related but exhibit a domain shift, resulting in varying distributions between them. Typically, source domain datasets are abundant in labeled data, while the target domain possesses limited or even no labeled data. The primary objective of domain adaptation research is to develop a model that performs effectively in both domains, leveraging knowledge acquired from the labeled source domain and applying it to the unlabeled or sparsely labeled target domain. This problem manifests in various forms, including semi-supervised, weakly-supervised, or unsupervised domain adaptation.

In semi-supervised and weakly-supervised settings, there is some amount of labeling information available for the target domain data, offering a

middle ground in terms of difficulty. On the other hand, the unsupervised setting poses the greatest challenge, as it lacks any labeling information for the target domain data. Successfully addressing this scenario requires the model to generalize effectively without relying on labeled target domain samples, making it the most demanding and intricate setting in domain adaptation research.

2.2.1 Unsupervised Domain Adaptive Object Detection

Unsupervised Domain Adaptive Object Detection represents a field within computer vision, addressing the inherent challenges of deploying object detection models across diverse and unlabelled domains. In scale of difficulty, adapting a model that performs well in a domain to an unseen environment that comes from a shifted distribution of data is a formidable problem to solve. These methods seek to close the gap between domains by leveraging knowledge from a labeled source domain and applying it to the unlabeled target domain. Several ways to tackle this problem and align both domains include the use of adversarial feature learning [7–9] and knowledge distillation with mean-teacher training [5, 10].

2.3 Edge Information

Edge information has been leveraged in many tasks including object detection [11–15]. [14] explored this concept in the field of medical image segmentation, utilizing a Canny edge detector to extract additional edge information from the data and paired it with an adversarial loss to gen-

erate invariant features. [15] also placed special attention to the edges in their Cycle GAN implementation for medical images segmentation, to retain these low-level details during the reconstruction step. Both of these methods show the usefulness of edge information in domain adaptation, however they both make use of the canny edge detector, that may not be enough in some more complex situations.

Incorporating edge information into our research stems from the observation that edges can be extracted from the shallow layers of a model, making them a valuable source of domain-invariant information. Notably, edges play a crucial role when dealing with domain shifts, as they often remain consistent across different domains. For instance, the edges of an object, like a car, remain relatively stable whether captured on a sunny or rainy day. That is why in our proposed pipeline, we decide to add edge information by using an auxiliary lightweight network.

Moreover, in the context of camouflaged object detection, [16] mentions that an observer, to find a camouflaged object, will first find region proposals for the objects, and then analyze these regions more closely to find the hidden objects. The OSFormer [17], employed a DETR for camouflaged object segmentation, and they found that weak boundary cues are essential for their task. They incorporated an edge module to the DETR pipeline, adding this rich low-level information to the model. We hypothesize that this same intuitions and ideas can be applied to domain adaptation, where the DETR model can also benefit by focusing on object-related details and paying less attention to background variations, treating the objects in the

target domain a camouflaged version of the ones in the source domain.

Chapter III
Methodology

This section introduces our Domain Adaptive Edge-guided DETR module for Object Detection (DAEG). As normal in the unsupervised domain adaptation, the input data includes the labeled source images and the unlabeled target images. Our goal is to increase the generability of the features learned by the DETR model so that it improves its performance on the target domain. As the module is easily adapted to the DETR architecture, we show its usefulness when paired with a SOTA-performing model.

3.1 Method overview

We start by introducing the problem setting and the based detector used for the experiments. In this context let us define the source dataset as $S = \{X_{is}, Y_{is}\}_{i=1}^{Ns}$, and it consists of Ns number of images. X_{is} will represent the i-th image and Y_{is} denotes the corresponding bounding box annotations with their category label. Similarly, let us denote the target dataset as $T = \{X_{it}\}_{i=1}^{Nt}$ having Nt number of target domain images with no ground-truth annotations. Since there is a noticeable change in the illumination and clarity of objects between the two domains, there is a clear domain shift that needs to be addressed.

To implement this work, we used the Deformable DETR as the base detector. As illustrated in Fig. 3.1, it is formed by a CNN backbone that extracts features, an encoder which enhances them, a decoder which interprets them and finally a feed-forward network that use them to make the final predictions. In this architecture, the features learned from the backbone are considered low-level, so enhancing them would improve the model's overall adaptability to new domains. With this objective in mind, we propose a novel module which incorporates information from the backbone, the encoder, and to enhance low-level features even more, an external auxiliary edge-detector network is also used as input.

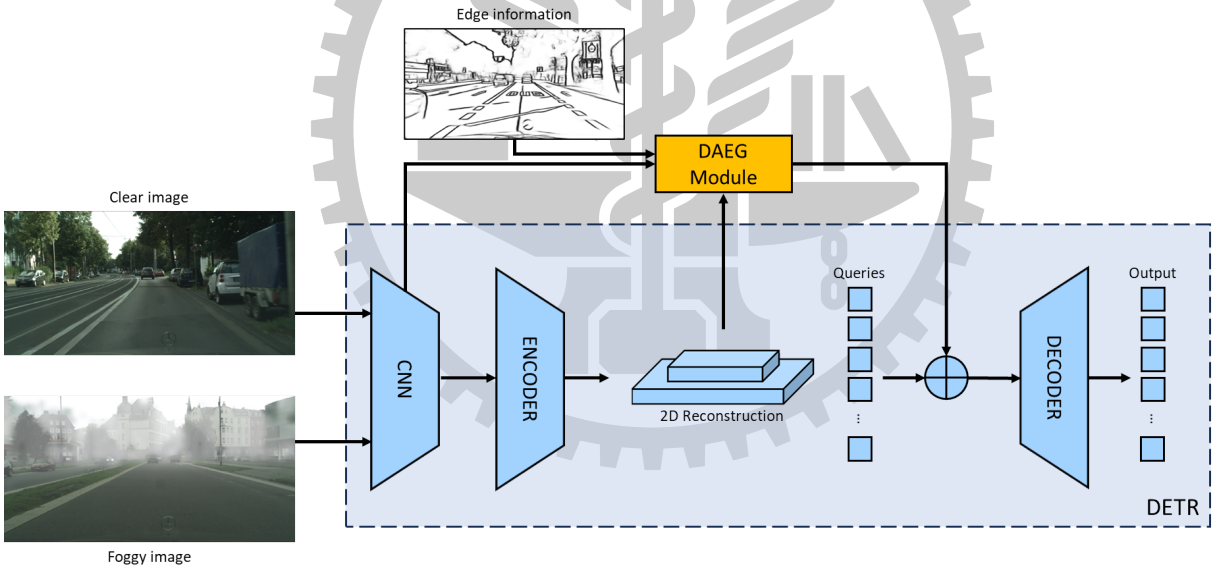


Figure 3.1: DETR architecture pipeline with the DAEG module.

3.1.1 Domain Adaptive Edge-guided DETR Module

The purpose of this addition, depicted in Fig. 3.2, is to enrich the decoder queries by integrating a combination of low and high-level information acquired from previous stages.

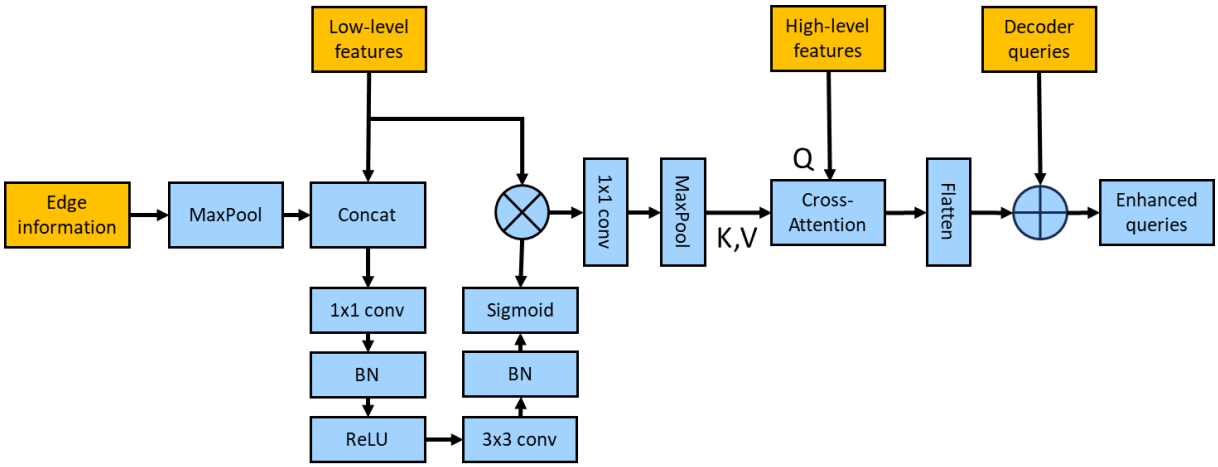


Figure 3.2: DAEG module structure.

This module receives inputs from three different sources, the backbone, the encoder, and an auxiliary lightweight network which extracts the edges from the input images. Firstly, the edge information and backbone inputs fusion together by following Eq. 1.

$$L' = L \otimes (3x3Conv(1x1Conv(Cat(MaxPool(E), L))) \tag{1}$$

In this equation, E refers to the edge information. L refers to the low-level features from the encoder. The notion Cat refers to the concatenation function. The backbone features are obtained in a 1/8 scale, while the higher-level features from the encoder are at a 1/32 scale of the original input image size. The 1x1 convolution acts as a channel shuffle to embed the edge information into the low-level features, to finally produce the enriched low-level features denoted by L'.

The process of edge extraction is carried out through the use of a lightweight network that has undergone fine-tuning to optimize its performance for the given input data. The emphasis here is on achieving utmost

clarity in edge delineation, a critical factor, particularly within the target domain where extracting edges poses increased difficulty. The decision to employ a sophisticated network, as opposed to a more straightforward algorithm such as the Canny Edge Detector, stems from the need to address the extra complexities inherent in the target domain's data. This strategic choice aims to ensure a more robust and precise edge extraction, even under challenging conditions, without introducing too much extra computation.

Subsequently, the higher-level features from the encoder and the newly generated low-level features undergo processing within a cross-attention module following Eq. 2.

$$\text{Cross-Attn:} \quad Q_q = H, K_{k,v} = MaxPool(1x1Conv(L')) \tag{2}$$

Here, H refers to the higher-level features from the encoder. This transformative process yields features that, upon reshaping, seamlessly merge with the original decoder queries, resulting in enhanced queries infused with essential low-level edge information. This meticulous integration ensures a holistic and enriched representation of the input, promoting improved performance in subsequent stages of the decoding process.

Chapter IV
Experiments

4.1 Dataset

The model will be trained using the fully labeled Cityscapes dataset [18]. Cityscapes is a comprehensive dataset containing high-resolution images from urban environments, making it suitable for training object detection models. An example image from the dataset is shown in Fig. 4.1. To evaluate the model's domain adaptation capabilities, the unlabeled Foggyscapes dataset will be used. Foggyscapes artificially adds fog to the Cityscapes data, which is an ideal scenario for testing the model's adaptability to different domains. A comparison of the two domains is presented in Fig. 4.2. Some key factors that make the Cityscapes dataset great for this task are:

- **Annotations:** The dataset offers detailed pixel-level annotations initially designed for semantic segmentation tasks. However, due to the inclusion of instance-level annotations, which meticulously outline individual objects with precise boundaries using bounding boxes, the dataset can be seamlessly adapted to accommodate object detection tasks. These annotations encompass a diverse array of object classes commonly encountered in urban scenes, thereby furnishing a compre-

hensive ground truth for a multitude of computer vision tasks.

- **Number of Classes:** Cityscapes includes a 30 distict classes. These classes contain a wide range of urban elements, including but not limited to pedestrians, vehicles, buildings, road signs, and vegetation.

- **Challenging Scenarios:** The dataset intentionally includes challenging scenarios such as occlusions, dynamic traffic situations, and complex urban layouts.

Figure 4.1: Example image from the Cityscapes dataset.

Figure 4.2: Comparison between the source (left) and target (right) domains.

4.2 Implementation Details

We evaluate the usefulness on the module by improving the O2Net [9], which adds an adversarial loss to the backbone and another alignment loss to the decoder. This method was utilized because it is the best-performing method that implements the DETR framework. The training of the Deformable DETR model still follow standard practices, including the selection of loss functions, optimization techniques, and training epochs. In addition, we adopt Mean Average Precision (mAP) with a threshold of 0.5 as the evaluation metric. All experiments were conducted on NVIDIA GeForce RTX 3090 GPUs.

4.3 Main Results

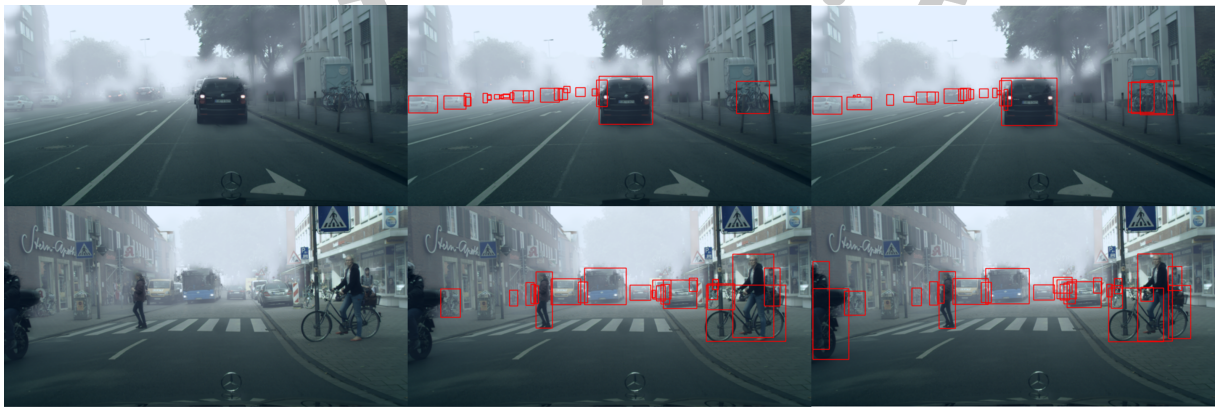
Diverse weather conditions are frequently encountered, presenting significant challenges that require adaptive responses from object detectors. In this study, we evaluate the effectiveness of our newly implemented module within the O2Net framework, comparing it with several other methods in Table 4.1. The results demonstrate the superior performance of our final model across various categories, enhancing the baseline O2Net's original accuracy from 46.8% to 48.7%.

In Fig. 4.3, we present visualizations of the results obtained with our model on the Foggy Cityscapes dataset, accompanied with the corresponding ground truth labels. In these cases, the performance of the model seems acceptable, and it is evident on these examples that the elements that were

Table 4.1: Object detection performance comparison in % on CS → Foggy CS. Baseline models are marked with an asterisk(*) sign. Best results marked in **bold**.

Method	Detector	Person	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
Faster RCNN*	Faster RCNN	26.9	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DivMatch [19]	Faster RCNN	31.8	51.0	20.9	41.8	34.3	26.6	32.4	34.9
SWDA [20]	Faster RCNN	31.8	48.9	21.0	43.8	28.0	28.9	35.8	35.3
SCDA [21]	Faster RCNN	33.8	52.1	26.8	42.5	26.5	29.2	34.5	35.9
MTOR [22]	Faster RCNN	30.6	44.0	21.9	38.6	40.6	28.3	35.6	35.1
CR-DA [23]	Faster RCNN	30.0	46.1	22.5	43.2	27.9	27.8	34.7	34.2
CR-SW [23]	Faster RCNN	34.1	53.5	24.4	44.8	38.1	26.8	34.9	37.6
GPA [24]	Faster RCNN	32.9	54.1	24.7	45.7	41.1	32.4	38.7	39.5
MIC (SADA) [25]	Faster RCNN	50.9	**67.0**	33.9	52.4	33.7	40.6	47.5	47.6
Deformable DETR*	D-DETR	38.0	45.3	16.3	26.7	4.2	22.9	36.7	28.6
SFA [26]	D-DETR	46.5	62.6	25.1	46.2	29.4	28.3	44.0	41.3
MTTrans [6]	D-DETR	47.7	65.2	25.8	49.5	33.8	32.6	46.5	43.4
O2net [9]	D-DETR	48.7	63.6	31.1	47.6	**47.8**	38.0	45.9	46.8
DAEG (O2Net) [9]	D-DETR	**51.9**	48.6	**47.7**	**53.3**	28.3	**67.9**	**54.9**	**48.7**

not correctly identified were either highly occluded or partially obscured, rendering them particularly challenging for natural detection. This difficulty extends not only to the model but also to human perception in some cases.



(a) Original Image (b) Our result (c) Ground truth

Figure 4.3: Qualitative results of the model, compared to the ground truth.

For the edge extractor network, the outputs are illustrated in Fig. 4.4. It is noticeable that the new added input represents a simplified version of the input images, great at capturing rich low-level information that our

module integrates and that the DETR leverages.

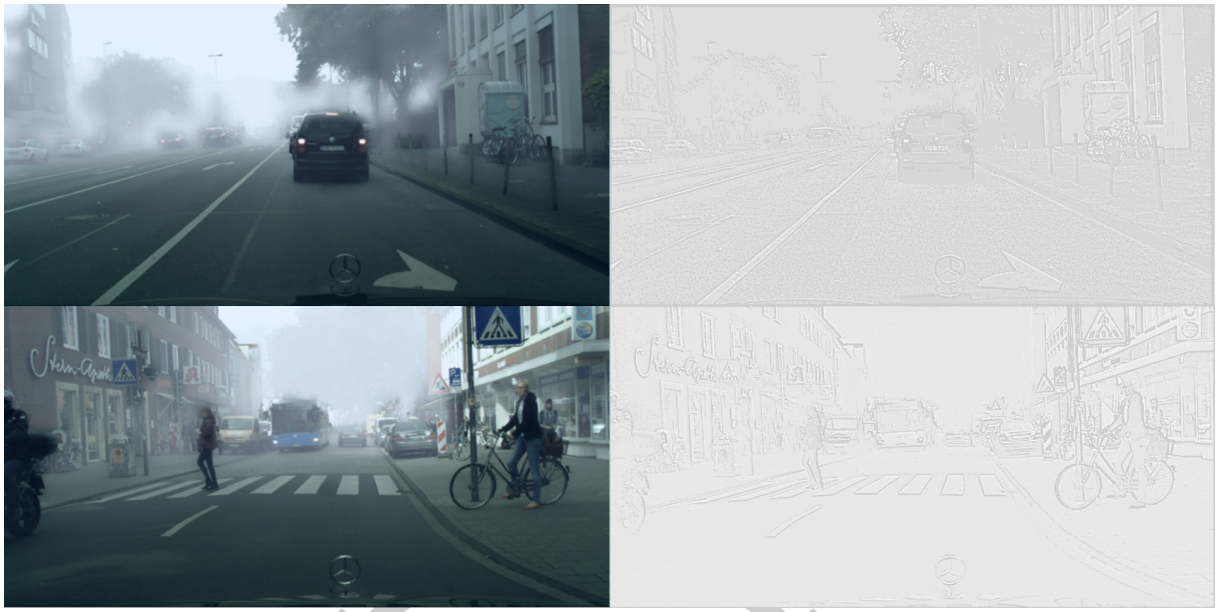


Figure 4.4: Output of the auxiliary edge detection network. Left: Input image. Right: Network output.

For an insight into the content transmitted to the queries, we can refer to Fig. 4.5 where it is evident that the cross-attention result effectively emphasizes the targeted areas, successfully enhancing the information incorporated into the decoder queries.

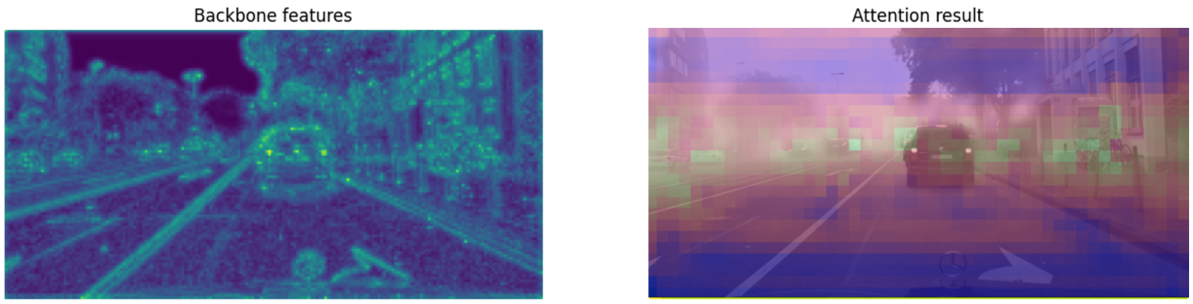


Figure 4.5: Information within the DAEG module. Left: Backbone information with enhanced edges. Right: Result of the cross-attention operation between high-level and lower-level information.

Furthermore, to demonstrate the efficacy of the module when integrated with the base Deformable DETR, Table 4.2 is presented. The model ex-

hibits an improvement from its base accuracy of 28.6% to 29.8%, substantiating the concept that our approach successfully enhances the generalization capacity of the model through the seamless integration of a plug-and-play module.

Table 4.2: Performance on the base Deformable DETR in % on CS → Foggy CS. Best results marked in **bold**.

Method	Detector	Person	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
Faster RCNN	Faster RCNN	26.9	35.6	**18.3**	32.4	9.6	25.8	28.6	26.9
Deformable DETR	D-DETR	38.0	**45.3**	16.3	26.7	4.2	22.9	**36.7**	28.6
DAEG (D-DETR)	D-DETR	**39.0**	41.7	2.5	**39.2**	**11.9**	**44.6**	25.9	**29.8**

Chapter V
Discussion

Initially, it is crucial to note that domain adaptation for object detection remains a vibrant area of research with considerable potential for growth. While numerous models exhibit performance around 70% on the original Cityscapes dataset [27, 28], the best performing methods in the domain adaptation field show a significant drop of approximately 20% when applied to the modified Foggyscapes dataset. This indicates the ongoing need for enhancements to achieve a domain-invariant model.

In the context of the results obtained in this study, we believe that the integration of edge information into the DETR structure has not been extensively explored in the realm of domain adaptation, thereby presenting a promising avenue for future research. Our findings demonstrate that the inclusion of this low-level information contributes to improving the generality of the models, enabling them to focus on features shared across all domains. Despite surpassing the baseline performance, further exploration of the incorporation of edge information remains possible. For instance, it could be investigated by directly integrating the information into the decoder layers or incorporating it within an adversarial loss, a strategy that has proven valuable in this field.

Ultimately, the trajectory of domain adaptation for object detection can be shaped by research dedicated to mitigating unreal assumptions inherent in our work. For instance, our current approach operates under the assumption of balanced class distribution between the source and target domains, as well as the presence of shared classes between the two domains. Addressing and refining these assumptions will be instrumental in advancing the robustness and applicability of domain adaptation methods in object detection.

Chapter VI
Conclusion

In conclusion, our research presents a novel plug-and-play module designed to augment the generalization capacity of the DETR framework. Through its incorporation, we achieved a noteworthy enhancement in the performance of a state-of-the-art competitor, elevating it from 46.8% to 48.7%, reflecting a notable 1.9% improvement. The introduced module effectively harnesses domain-agnostic information, contributing to the refinement of decoder queries. This, in turn, enhances the overall quality of the detection process, showcasing the potential of our approach in advancing the capabilities of the DETR model.

Bibliography

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[4] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[5] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," in *CVPR*, 2021, pp. 4091–4101.

[6] J. Yu, J. Liu, X. Wei, H. Zhou, Y. Nakata, D. Gudovskiy, T. Okuno, J. Li, K. Keutzer, and S. Zhang, "Mttrans: Cross-domain object detection with mean teacher transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 629–645.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[8] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3339–3348.

[9] K. Gong, S. Li, S. Li, R. Zhang, C. H. Liu, and Q. Chen, "Improving transferability for domain adaptive detection transformers," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1543–1551.

[10] J. Yu, J. Liu, X. Wei, H. Zhou, Y. Nakata, D. Gudovskiy, T. Okuno, J. Li, K. Keutzer, and S. Zhang, "Cross-domain object detection with mean-teacher transformer," in *ECCV*, 2022.

[11] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7264–7273.

[12] E. Q. Wu, X. Peng, C. Z. Zhang, J. Lin, and R. S. Sheng, "Pilots'' fatigue status recognition using deep contractive autoencoder network," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 10, pp. 3907–3919, 2019.

[13] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1623–1632.

[14] W. Yan, Y. Wang, M. Xia, and Q. Tao, "Edge-guided output adaptor: Highly efficient adaptation module for cross-vendor medical image segmentation," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1593–1597, 2019.

[15] T. Vo and N. Khan, "Edge-preserving domain adaptation for semantic segmentation of medical images," *arXiv preprint arXiv:2111.09847*, 2021.

[16] G.-P. Ji, L. Zhu, M. Zhuge, and K. Fu, "Fast camouflaged object detection via edge-based reversible re-calibration network," *Pattern Recognition*, vol. 123, pp. 108414, 2022.

[17] J. Pei, T. Cheng, D.-P. Fan, H. Tang, C. Chen, and L. Van Gool, "Osformer: One-stage camouflaged instance segmentation with transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 19–37.

[18] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12456–12465.

[20] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[21] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 687–696.

[22] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11457–11466.

[23] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11724–11733.

[24] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12355–12364.

[25] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11721–11732.

[26] W. Wang, Y. Cao, J. Zhang, F. He, Z.-J. Zha, Y. Wen, and D. Tao, "Exploring sequence feature alignment for domain adaptive detection transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1730–1738.

[27] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022.

[28] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al., "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14408–14419.