ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

CROSS-DOMAIN ONE-SHOT OBJECT DETECTION BY ONLINE FINE-TUNING

M.Sc. THESIS

İrem Beyza ONUR

Department of Electronics and Communication Engineering

Telecommunication Engineering Programme

JUNE 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

CROSS-DOMAIN ONE-SHOT OBJECT DETECTION BY ONLINE FINE-TUNING

M.Sc. THESIS

İrem Beyza ONUR
(504211318)

Department of Electronics and Communication Engineering

Telecommunication Engineering Programme

Thesis Advisor: Prof. Dr. Bilge GÜNSEL

JUNE 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ÇEVRİMİÇİ İNCE-AYAR İLE TEK-ÖRNEKLİ ÇAPRAZ-ALAN NESNE TESPİTİ

YÜKSEK LİSANS TEZİ

İrem Beyza ONUR
(504211318)

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Telekomünikasyon Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Bilge GÜNSEL

HAZİRAN 2024

İrem Beyza ONUR, a M.Sc. student of ITU Graduate School student ID 504211318 successfully defended the thesis entitled "CROSS-DOMAIN ONE-SHOT OBJECT DETECTION BY ONLINE FINE-TUNING", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Bilge GÜNSEL** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Cihan TOPAL** Istanbul Technical University	.....
	Asst. Prof. İnci Meliha BAYTAŞ Boğaziçi University	.....

Date of Submission : **22 May 2024**
Date of Defense : **26 June 2024**

To my family,

FOREWORD

First and foremost, I would like to present my sincere gratitude to my advisor, Prof. Dr. Bilge GÜNSEL, for their continuous support, patience, motivation, and immense knowledge. Over the past three years, her guidance has been invaluable throughout my research and the writing of this thesis.
I would also like to extend my heartfelt thanks to TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA), for providing the computational resources necessary for the numerical calculations reported in this thesis. Their support has been instrumental in the completion of this research.
I would like to express my endless thanks to my family. They gave me constant support and compassion during the toughest times. Their encouragement has been crucial to my perseverance in this research, and I am immensely grateful to have them.
Finally, I would like to thank my friends for their continuous support and for making difficult times easier to get through.

June 2024

İrem Beyza ONUR
(Electronics and Communication Engineer)

TABLE OF CONTENTS

ABBREVIATIONS

OSOD	**:** One-shot Object Detection
FSOD	**:** Few-shot Object Detection
OSCDA	**:** One-shot Cross-domain Adaptation
SACDA	**:** Shot Augmentation for Cross-domain Adaptation
CDQSS	**:** Cross-domain Query Shot Selection
IoU	**:** Intersection over Union
CD-OSL	**:** Cross-domain One-shot Learning
CD-FSOD	**:** Cross-domain Few-shot Object Detection
FPN	**:** Feature Pyramid Networks
RPN	**:** Region Proposal Network
BCE	**:** Binary Cross Entropy
AP	**:** Average Precision

SYMBOLS

$\boldsymbol{Q}$	: Query Shot
$\boldsymbol{T}$	: Target Image
$\boldsymbol{N}$	: Novel Class
$\boldsymbol{B}$	: Base Class
$\mathbf{W_S}$	: Spatial Aware Weight
$\mathbf{F_T}$	: Feature Map of Target Image
$\mathbf{F_Q}$	: Feature Map of Query Shot
$\mathbf{R_C}$	: Contrastive Level Relation
$\mathbf{R_S}$	: Salient Level Relation
$\mathbf{R_A}$	: Attention Level Relation
$\mathbf{R}$	: Fused Relation
$\mathbf{D(.)}$	: Depthwise Convolution
$\mathbf{P(.)}$	: Pointwise Convolution
$\mathbf{L_{cls}}$	: Classification Loss
$\mathbf{L_{reg}}$	: Regression Loss
$\mathbf{L_{total}}$	: Total Loss
$\mathbf{P}$	: Precision
$\mathbf{R}$	: Recall
mAP	: Mean Average Precision
nAP	: Mean Average Precision for Novel Classes

LIST OF TABLES

Page

LIST OF FIGURES

CROSS-DOMAIN ONE-SHOT OBJECT DETECTION BY ONLINE FINE-TUNING

SUMMARY

Object detection aims to identify and locate objects within an image or video frame. Recently, deep learning-based object detectors have made significant contributions to the community, thanks to their capability of processing and learning from large volumes of data. However, their detection performance heavily depends on large labeled datasets, which are essential for them to generalize effectively across previously unseen (novel) or rare classes. This restriction is addressed by the recent development of few-shot object detection (FSOD) and one-shot object detection (OSOD) techniques, which aim to detect novel classes using a few or a single sample of a previously unseen class, enabling rapid adaptation to novel classes without extensive labeled data. The FSOD and OSOD paradigms which allow models to adapt to novel classes from a small sample size have two main lines: methods based on meta-learning and those based on transfer learning. Meta-learning approaches aim to learn a generalizable set of parameters on data-abundant base classes by applying an episodic training approach. This approach is designed to transfer the base knowledge gained on data-abundant base classes to data-scarce novel classes during the inference phase. The other methods based on transfer learning focus on fine-tuning the model, which has trained on extensive datasets, on the limited number of new examples by applying several methods such as freezing only the specific layers or adaptive learning rate scheduling during the fine-tuning. In FSOD and OSOD techniques, the model is trained on data-abundant base classes and then fine-tuned on both the base and novel classes or on only the novel classes. This methodology makes them well-suited approaches for use with still images where the task is to recognize objects based on limited data.

Although FSOD and OSOD methods enable quick adaptation to novel objects, their performance highly depends on the training domain, typically composed of still images. Due to the domain shift, the conventional setup yields significant performance degradation in one-shot or few-shot object detection in cross-domain evaluations. Moreover, most of the recent studies remain limited to focusing on performance gaps between different image domains, rather than those between image and video domains.

Differing from the existing work, this thesis particularly focuses on the significant performance gap observed in cross-domain evaluations, from the still image domain to the video domain, which has been largely overlooked in the literature. In video domain evaluations, the main purpose of the detection model is to detect the target object, which is introduced at the beginning of the video sequence, within subsequent frames successfully. In the scope of this thesis, we choose to work with the OSOD

models rather than the FSOD models because object detection in video sequences primarily hinges on the model's ability to adapt to the target object using only a single example. Therefore, OSOD models, designed to adapt to the target object using only one example, are more suitable than FSOD models, which necessitate few examples to adapt to the target object. In particular, OSOD models aim to classify and localize a target object in an image using its particular representation known as a query shot. This is achieved through a template-matching algorithm that detects all the instances from the class of this single query shot within the target image. The paradigm enables the model to adapt to the specific appearance of the target object from a single sample, in contrast to FSOD where the model is designed to adapt to novel classes using a small number of samples.

In addition to the scarcity of examples of the target object aimed to be detected throughout the video frames, temporal challenges, and motion changes also present a substantial challenge for OSOD methods. OSOD models' ability to handle affine motion changes and maintain temporal consistency is crucial for sufficient performance in video sequences. Moreover, the challenge OSOD brings along arises from the requirement that the model must adapt to novel classes based solely on a single sample, which demands a higher level of generalization and adaptability. The majority of the OSOD models are trained and evaluated on the same domain in which the data distributions and class characteristics are quite similar between the training and evaluation sets. Although recent studies have examined the cross-domain evaluation of OSOD models, they have primarily focused on evaluations within different still-image domains, rather than between still-image and video domains. The models are vulnerable to severe shifts in data distribution due to the domain shift between still-image and video domains.

In this thesis, we aim to demonstrate and analyze the reasons behind the performance gap that OSOD models experience in cross-domain scenarios. To do this, we evaluate a state-of-the-art (SOTA) OSOD model, BHRL, which has been trained on the MS COCO dataset from the still-image domain, using the VOT-LT 2019 dataset, which presents the challenging context of the video domain. For a fair evaluation, we include only the video frames where the target object is present. To alleviate the performance degradation, we take BHRL as the baseline OSOD model and propose three different novel OSOD frameworks, based on integrating an online fine-tuning scheme and a query shot update mechanism into the inference architecture of BHRL, the SOTA OSOD model utilizing multi-level feature learning. During the evaluation of the proposed frameworks, the mAP0.5 metric is used and class-based reporting is performed. Classes included in the training phase were classified as base classes, while those that were not included were categorized as novel classes. In the following, the proposed OSOD frameworks are summarised.

1. **OSCDA w/o CDQSS:** The initial appearance of the target object is taken as the query shot and the model is online fine-tuned only on this query shot once at the beginning of the inference phase. Subsequently, the model tries to detect all instances of the relevant class within the video frames. Although fine-tuning is a conventional approach in transfer learning, integration with the multi-level feature

learning of BHRL improves the detection mAP.50 performance by 14% on all classes.

2. **OSCDA w/ CDQSS:** A major limitation of OSCDA w/o CDQSS in video object detection is its vulnerability to rapid appearance changes of the target object, resulting from the risk of overfitting on the query shot that represents the target's initial appearance. To overcome this drawback, in addition to fine-tuning, CDQSS, which is an adaptive query shot selection module, is integrated into the baseline architecture. This approach enables unsupervised online fine-tuning to deal with the rapid appearance changes of the target object caused by the affine motion throughout the video frames. By CDQSS, the query shot is updated with the model's detections based on their objectness scores and localization consistency across frames. The model is continuously fine-tuned with the query shots chosen by the CDQSS during the inference phase. The fine-tuning process is called unsupervised fine-tuning since it is based solely on the model's detections rather than the ground truth. CDQSS provided an additional 6% improvement in mAP.50 on all classes.

3. **SACDA:** In order to take advantage of extra shots without leaving the one-shot detection approach, we propose incorporating online fine-tuning into the BHRL using the initial query shot and its synthetically generated variations referred to as augmented shots. In particular, similar to OSCDA w/o CDQSS, SACDA conducts fine-tuning only once at the beginning of the inference, and then the model tries to detect all instances of the relevant class throughout the subsequent frames. SACDA aims to adapt to quick changes in the target object's appearance, such as flipped or rotated versions, without relying on the continuous fine-tuning process suggested in the second framework (OSCDA w/ CDQSS). SACDA improves BHRL's mAP.50 score by 14%, matching the improvement seen with OSCDA (w/o CDQSS). However, SACDA significantly outperforms the previous frameworks in specific sequences such as ballet, group2, and longboard by 14%, 28%, and 46% respectively. These sequences share challenges such as extreme changes in lighting, scale, and rotation of target objects, as well as rapid illumination changes. Given SACDA's design goal to enhance the OSOD model's robustness to variations in target and scene appearances, these significant gains indicate SACDA's potential effectiveness.

The achieved performance improvements demonstrate the proposed methods' effectiveness in tackling domain shift challenges faced during the cross-domain evaluations for video object detection.

ÇEVRİMİÇİ İNCE-AYAR İLE TEK-ÖRNEKLİ ÇAPRAZ-ALAN NESNE TESPİTİ

ÖZET

Nesne tespiti, bir görüntü veya video karesi içindeki nesneleri sınıflamayı ve yerlerini belirlemeyi amaçlar. Son zamanlarda, büyük veri miktarlarını işleyebilme ve bu verilerden öğrenebilme kapasiteleri sayesinde derin öğrenme tabanlı nesne tespit algoritmaları literatüre önemli katkılarda bulunmuştur. Ancak, bu algoritmaların tespit performansları, genellikle görülmemiş (yeni) veya nadir sınıflar arasında etkili bir şekilde genelleme yapabilmeleri için büyük boyutlu ve etiketli veri kümelerine büyük ölçüde bağımlıdır. Bu kısıtlama, az sayıda veya tek bir örnek kullanarak yeni sınıfları tespit etmeyi amaçlayan az-örnekli nesne tespiti (FSOD) ve tek-örnekli nesne tespiti (OSOD) tekniklerinin son gelişmeleri ile ele alınmaktadır. Bu paradigmalarda modellerin küçük örnek boyutlarından yeni sınıflara adapte olabilmesi, meta-öğrenme ve aktarım öğrenimi tabanlı yöntemler olmak üzere iki ana hat üzerinde durulmuştur. Meta-öğrenme yaklaşımları, episodik eğitim uygulayarak veri-bolluğu olan baz sınıflar üzerinde genellenebilir bir parametre seti öğrenmeyi amaçlar. Bu yaklaşım, çıkarım aşamasında veri-yoksunu yeni sınıflara baz sınıf bilgisinin aktarılmasını sağlamak için tasarlanmıştır. Aktarım öğrenimi tabanlı diğer yöntemler ise, geniş veri setleri üzerinde eğitilmiş modele, yeni örneklerin sınırlı sayıda olması gibi durumlarla başa çıkmak için, belirli katmanların dondurulması veya uyarlamalı öğrenme hızı programlaması gibi çeşitli yöntemler uygulayarak ince ayar yapmaya odaklanır. FSOD ve OSOD tekniklerinde, model öncelikle veri-bolluğu olan baz sınıflar üzerinde eğitilir ve sonra modele hem baz hem de yeni sınıflar üzerinde veya sadece yeni sınıflar üzerinde ince ayar yapılır. Bu metodoloji, özellikle sınırlı veriye dayalı nesne tanıma görevleri için uygun yaklaşımlar sunar.

FSOD ve OSOD yöntemleri yeni nesnelere hızlı uyum sağlamayı mümkün kılmasına rağmen, performansları büyük ölçüde genellikle durağan görüntülerden oluşan eğitim alanına bağlıdır. Alan değişimi nedeniyle, geleneksel kurulum, çapraz alan değerlendirmelerinde tek örnekli veya az örnekli nesne algılamada önemli performans düşüşlerine yol açar. Ayrıca, son çalışmaların çoğu, durağan görüntü alanları arasındaki performans farklarına odaklanmakla sınırlı kalmakta ve görüntü ile video alanları arasındaki farklara yeterince dikkat etmemektedir.

Mevcut çalışmalardan farklı olarak, bu tez özellikle literatürde büyük ölçüde göz ardı edilen, durağan görüntü alanından video alanına geçişte gözlemlenen önemli performans farkına odaklanmaktadır. Video alanındaki değerlendirmelerde, algılama modelinin ana amacı, video dizisinin başında tanıtılan hedef nesneyi sonraki karelerde başarılı bir şekilde tespit etmektir. Bu tezin kapsamında, video dizilerinde nesne algılama esas olarak modelin yalnızca tek bir örnek kullanarak hedef nesneye uyum

sağlama yeteneğine bağlı olduğundan, FSOD modelleri yerine OSOD modelleri ile çalışmayı tercih ediyoruz. Bu nedenle, yalnızca bir örnek kullanarak hedef nesneye uyum sağlamayı amaçlayan OSOD modelleri, hedef nesneye uyum sağlamak için birkaç örneğe ihtiyaç duyan FSOD modellerine kıyasla daha uygundur. OSOD modelleri, bir sorgu çekimi olarak bilinen belirli bir temsili kullanarak bir görüntüdeki hedef nesneyi sınıflandırmayı ve yerelleştirmeyi amaçlar. Bu, tek bir sorgu çekiminden sınıfın tüm örneklerini hedef görüntüde tespit eden bir şablon eşleme algoritması aracılığıyla gerçekleştirilir. Bu paradigma, modelin tek bir örnekten hedef nesnenin belirli görünümüne uyum sağlamasını mümkün kılar; oysa FSOD'de model, yeni sınıflara küçük bir örnek sayısı kullanarak uyum sağlamak üzere tasarlanmıştır.

Hedef nesnenin video kareleri boyunca tespit edilmesi amacıyla örneklerin azlığına ek olarak, zamansal zorluklar ve hareket değişiklikleri de OSOD yöntemleri için önemli bir zorluk teşkil etmektedir. OSOD modellerinin afin hareket değişiklikleriyle başa çıkma ve zamansal tutarlılığı koruma yetenekleri, video dizilerinde yeterli performans için hayati öneme sahiptir. Ayrıca, OSOD'nin beraberinde getirdiği zorluk, modelin yalnızca tek bir örneğe dayanarak yeni sınıflara uyum sağlaması gerektiği gerçeğinden kaynaklanmaktadır. Bu durum, daha yüksek düzeyde genelleme ve uyum sağlama yeteneği gerektirir. OSOD modellerinin çoğu, veri dağılımlarının ve sınıf özelliklerinin eğitim ve değerlendirme setleri arasında oldukça benzer olduğu aynı alanda eğitilmekte ve değerlendirilmektedir. Son çalışmalar, OSOD modellerinin çapraz alan değerlendirmesini incelemiş olmasına rağmen, genellikle farklı durağan görüntü alanları içindeki değerlendirmelere odaklanmıştır, durağan görüntü ve video alanları arasındaki değerlendirmelere değil. Modeller, durağan görüntü ve video alanları arasındaki alan değişimi nedeniyle ciddi veri dağılımı değişimlerine karşı savunmasızdır.

Bu tezde, OSOD modellerinin çapraz alan senaryolarında karşılaştıkları performans farkının nedenlerini göstermek ve analiz etmek amaçlanmaktadır. Bunu yapmak için, durağan görüntü alanından MS COCO veri seti üzerinde eğitilmiş olan son teknoloji bir OSOD modeli olan BHRL'yi, video alanının zorlu bağlamını sunan VOT-LT 2019 veri seti kullanarak değerlendiriyoruz. Adil bir değerlendirme için, yalnızca hedef nesnenin bulunduğu video karelerini dahil ediyoruz. Performans düşüşünü hafifletmek amacıyla, BHRL'yi temel OSOD modeli olarak alıp, BHRL'nin çok düzeyli özellik öğrenimini kullanan son teknoloji OSOD modelinin çıkarım mimarisine çevrimiçi ince ayar şeması ve sorgu çekimi güncelleme mekanizmasını entegre ederek üç farklı yeni OSOD çerçevesi öneriyoruz. Önerilen çerçevelerin değerlendirilmesi sırasında mAP0.5 metriği kullanılmış ve sınıf bazlı raporlama yapılmıştır. Eğitim aşamasında dahil edilen sınıflar temel sınıflar olarak sınıflandırılırken, dahil edilmeyenler yeni sınıflar olarak kategorize edilmiştir.
Aşağıda, önerilen OSOD çerçeveleri özetlenmiştir.

1. **OSCDA w/o CDQSS:** Hedef nesnenin ilk görünümü sorgu çekimi olarak alınır ve model, çıkarım aşamasının başında yalnızca bu sorgu çekimi üzerinde çevrimiçi olarak ince ayar yapılır. Daha sonra model, ilgili sınıfın tüm örneklerini video kareleri içinde tespit etmeye çalışır. İnce ayar yapmak transfer öğrenmede

geleneksel bir yaklaşım olsa da, BHRL'nin çok düzeyli özellik öğrenimi ile entegrasyon, tüm sınıflarda tespit mAP.50 performansını 14% artırmaktadır.

2. **OSCDA w/ CDQSS:** OSCDA w/o CDQSS'nin video nesne tespitindeki önemli bir sınırlaması, hedef nesnenin başlangıç görünümünü temsil eden sorgu çekimine aşırı uyum sağlama riskinden kaynaklanan hızlı görünüm değişikliklerine karşı savunmasız olmasıdır. Bu dezavantajı aşmak için, ince ayar yapmanın yanı sıra, uyarlanabilir bir sorgu çekimi seçme modülü olan CDQSS, temel mimariye entegre edilmiştir. Bu yaklaşım, video kareleri boyunca afin hareket nedeniyle meydana gelen hedef nesnenin hızlı görünüm değişiklikleriyle başa çıkmak için denetimsiz çevrimiçi ince ayar yapılmasını sağlar. CDQSS ile sorgu çekimi, nesnelilik skorları ve kareler arasındaki yerelleştirme tutarlılığına dayalı olarak modelin tespitleriyle güncellenir. Model, çıkarım aşamasında CDQSS tarafından seçilen sorgu çekimleriyle sürekli olarak ince ayar yapılır. İnce ayar süreci, zemin gerçeği yerine yalnızca modelin tespitlerine dayandığı için denetimsiz ince ayar olarak adlandırılır. CDQSS, tüm sınıflarda mAP.50'de ek olarak 6%'lık bir iyileşme sağlamıştır.
3. **SACDA:** Ek çekimlerden yararlanarak tek çekim algılama yaklaşımını terk etmemek için, BHRL'ye çevrimiçi ince ayar yapmayı ve başlangıç sorgu çekimi ile sentetik olarak oluşturulmuş varyasyonları (çoğaltılmış çekimler olarak adlandırılır) kullanmayı öneriyoruz. Özellikle, OSCDA w/o CDQSS'ye benzer şekilde, SACDA çıkarımın başında yalnızca bir kez ince ayar yapar ve ardından model, sonraki kareler boyunca ilgili sınıfın tüm örneklerini tespit etmeye çalışır. SACDA, hedef nesnenin görünümündeki hızlı değişimlere, örneğin ters veya döndürülmüş versiyonlara, ikinci çerçevede (OSCDA w/ CDQSS) önerilen sürekli ince ayar sürecine güvenmeden uyum sağlamayı amaçlar. SACDA, BHRL'nin mAP.50 skorunu 14% oranında artırarak, OSCDA (w/o CDQSS) ile görülen iyileşmeyle eşleşir. Ancak SACDA, bale, grup2 ve longboard gibi belirli dizilerde sırasıyla 14%, 28% ve 46% oranında önemli ölçüde daha iyi performans gösterir. Bu diziler, hedef nesnelerin aşırı aydınlatma, ölçek ve dönme değişiklikleri ile hızlı aydınlatma değişiklikleri gibi zorlukları paylaşır. SACDA'nın, hedef ve sahne görünümlerindeki değişikliklere karşı OSOD modelinin dayanıklılığını artırma tasarım hedefi göz önüne alındığında, bu önemli kazanımlar SACDA'nın potansiyel etkinliğini göstermektedir.

Önerilen çerçeveleri değerlendirmeden önce, baz alınan BHRL modeli kullanılarak OSOD modellerinin çapraz alan performans değerlendirmeleri sırasında deneyimlediği performans düşüşü gösterilmiştir. Bunun için hem MS COCO hem de Pascal VOC datasetleri ile eğitilmiş BHRL modeli kullanılmış, ve her iki model de hem kendi eğitim verisetinde hem de çapraz alandan alınan video verisetinde (VOT-LT 2019) değerlendirilmiştir. Sonuçlar göstermektedir ki, Pascal VOC veri seti ile eğitilen BHRL modelinin mAP0.5 performansı, kendi eğitim setinde yapılan değerlendirmelerle karşılaştırıldığında, video verisetinde tüm sınıflar, baz sınıflar ve yeni sınıflar için sırasıyla 34%, 30% ve 51% düşüş göstermektedir. Bu son derece dramatik düşüş, bu tezin ana motivasyonu olan OSOD modellerinin çapraz alan değerlendirmelerine karşı dayanıklı hale getirilmesi amacının dayanak noktası olarak düşünülmelidir.

Çapraz alan değerlendirmeleri sırasında OSOD modeli BHRL'nin performans düşüşü gösterildikten sonra önerilen çerçeveler çapraz video alanından alınan VOT-LT 2019 verisetinde değerlendirilmiştir. Önerilen çerçeveler, BHRL'nin tüm sınıflar arasındaki mAP.05 performansını sırasıyla 14%, 20% ve 14% oranında artırır. Sadece baz sınıflar üzerinde yapılan değerlendirmede ise, önerilen çerçeveler BHRL'nin mAP0.5 performansını sırasıyla 13%, 17% ve 13% oranında artırır. Bunun yanısıra, sadece yeni sınıflar üzerinde yapılan değerlendirmede ise BHRL'nin mAP0.5 artışı sırasıyla 17%, 28% ve 14% olarak gözlemlenmiştir. Belirtilmelidir ki, çapraz alan değerlendirmeleri için kullanılan video veriseti VOT-LT 2019, hedef objenin sahnede bulunmadığı kareler içermektedir. Bu kareler, literatürdeki tek atışlı nesne tespiti algoritmalarının değerlendirme yöntemleri göz önüne alınarak, adil bir karşılaştırma amacı ile göz ardı edilmiştir.

Sonuçlar ortaya koymaktadır ki, çapraz alan değerlendirmelerinde en yüksek mAP0.5 skorlarına ulaşan OSCDA (w/ CDQSS) çerçevesi olmuştur. Sınıflar bazında değerlendirmeler göz önüne alındığında, en fazla artış (28%) yeni sınıflar üzerinde yapılan değerlendirmeler sonucunda gözlemlenmiştir. Tek atışlı nesne algılama (OSOD) algoritmalarının ana motivasyonunun eğitim setinde görülmemiş veri sınıflarına tek bir örnek yardımıyla adaptasyon sağlanması olduğunu düşündüğümüzde bu sonuçlar umut vadetmektedir. Öte yandan, spesifik video dizilerinde önerilen üçüncü çerçevenin (SACDA) en yüksek mAP0.5 skoruna ulaştığı gözlemlenmektedir. Bu videolar incelendiğinde, video kareleri boyunca gerçeklerşen ve nesne tespit performansını düşüren birtakım zorlukların ortak olarak görüldüğü gözlemlenmiştir. Bu zorluklara örnek olarak, ani hedef görünüm değişiklikler, hızlı arka plan ve aydınlatma değişiklikleri ve ani kamera hareketleri verilebilir. Bu göstermektedir ki çeşitli veri arttırma yöntemleri ile sorgu atışının türevlerini üretmek ve model bu türevleri kullanarak ince ayar yapmak, spesifik videolarda başarılı sonuç vermiştir.

Gözlemlenen performans iyileştirmeleri, önerilen yöntemlerin çapraz alan değerlendirmelerinde performans düşüşüne neden olan alan değişikliği zorluklarıyla başa çıkma konusuundaki performansını göstermektedir.

1. INTRODUCTION

Object detection is a fundamental research area in computer vision, which aims to classify and localize objects of interest within an image. Recently, significant progress in deep neural networks led the community to build deep object detectors that have achieved state-of-the-art (SOTA) performance in object detection. However, deep object detectors require extensive labeled datasets to achieve good generalization performance across novel (previously unseen) classes, leading to a labor-intensive and expensive dataset preparation process. This restriction is addressed by the recent development of few-shot object detection (FSOD) techniques, which can be defined as the ability to generalize from only a few examples [5], [6], [7], [8], [9]. The FSOD methods aim to transfer the knowledge gained on data-abundant base classes in the training phase to the data-scarce novel classes in the inference phase.

One-shot object detection (OSOD) is the specific area of the FSOD, aiming to detect all instances of a novel class in a target image using a single query of a previously unseen class. Most one-shot object detectors focus on building semantic relations between the query and target pair with a template-matching algorithm and similarity learning [1], [10]. These approaches provide impressive performance in adaptation to novel classes using only a single query, however, their ability to generalize is limited since they focus on single-level feature relation between the query and target.

However, the main challenge one-shot object detection brings along arises from the requirement that the model must adapt to novel classes based solely on a single example, which demands a higher level of generalization and adaptability. To address this challenge, some of the recent developments [11], [12], [13] focus on building more complex and multi-level feature relations between the query and target, which help the model to learn the novel class representation at multiple levels. AIT [11] leverages the encoding capability of transformers to build a more complex and robust feature-based relation between the query and target. [12] utilizes different relation heads to establish semantic relations between query and target at multiple levels. BHRL [13], a baseline

model for our study, also leverages the advantages of multi-level feature learning, improving upon [12] by adding a fusion relation layer.

The majority of the one-shot and few-shot object detectors are trained and evaluated on the same domain in which the data distributions and class characteristics are quite similar between the training and evaluation sets. Baseline study BHRL [13] also uses the same domain for training and evaluation and outperforms other SOTA one-shot object detectors by utilizing multi-level feature learning. However, its performance highly relies on its training domain which comprises still images, resulting in a performance degradation caused by the challenges raised by domain shift between the still image and temporal video domain [14].

To tackle this problem, we propose to integrate an online one-shot fine-tuning layer into the inference architecture of BHRL [13]. In the inference phase, the model trained on the still image data from the COCO [15] is fine-tuned on the challenging VOT-LT 2019 [16] video data unseen in the training. This approach enables the domain adaptation (from still image domain to video domain) that most one-shot object detectors in the literature are not capable of. Unlike most one-shot and few-shot object detectors, we fine-tune the model on only the novel classes instead of a balanced set containing both novel and base classes. In particular, the ground truth of the target object in the first frame is assigned as the query shot for the initial fine-tuning, for each video sequence. To deal with challenges due to the rapid appearance changes of the target object across the video sequence, the query shot is updated by monitoring the classification scores as well as the localization consistency of the detections. The chosen detections of the model are used for the query update and fine-tuning, yielding unsupervised fine-tuning. Our contributions are summarized as follows:

- We demonstrate the performance degradation that arises from the domain shift by evaluating the performance of BHRL [13] on the VOT-LT video data [16]. Numerical results illustrate a 34% and 7% decrease in mAP0.5 obtained on the VOT-LT video data [16], using the BHRL model pre-trained on COCO [15] and Pascal VOC [17] datasets, respectively.

- We propose to integrate an unsupervised online fine-tuning scheme with the BHRL [13] to enable cross-domain adaptation in the OSOD task. Differing from the existing fine-tuning schemes, we designed an unsupervised mechanism that updates the target shot across the video frames. Performance tests on challenging VOT-LT video data set [16] demonstrate that the proposed method enhances the mAP0.5 of the baseline model by 18% and 28% on novel and base classes, respectively.

- Exploiting the augmentation by generating extra shots from the query shot, we demonstrate the impact of fine-tuning with transformed shots. An average increase of 14% in mAP0.5 of the baseline is reported across all classes on VOT-LT video sequences [16].

1.1 Literature Review

This section provides a brief literature overview of the key points of this thesis, FSOD and OSOD studies,cross-domain adaptation, and the motivation of this thesis.

1.1.1 Few-shot object detection (FSOD)

Most transfer learning-based FSOD methods [5], [6], [7], [8] use a two-stage fine-tuning approach (TFA) [5] for transferring the semantic information learned from only the abundant base classes to data-scarce novel classes. TFA [5] only fine-tunes the last layers of the ROI head, solely the box classifier and regressor, while freezing the rest of the network during the fine-tuning performed on a balanced dataset comprising of both base and novel classes. This approach is motivated by the desire to avoid the risk of overfitting the novel classes during the fine-tuning stage by preserving the knowledge gained from the base classes in the base training stage. TFA is the first study that introduces the fine-tuning approach for few-shot object detection and outperforms the earlier meta-learning-based methods.

The MPSR [6] uses a different fine-tuning strategy compared to TFA keeping all the network layers trainable during the fine-tuning stage. MPSR focuses on eliminating the scale scarcity in few-shot datasets introducing the multi-scale positive sample refinement branch. While this method is promising to alleviate the problem of imbalance between the number of positive and negative samples during the training,

its reliance on manual selection within the positive refinement branch may limit its generalization capacity.

FSCE [7] introduces the contrastive-aware object proposals to eliminate the risk of misclassifying the novel classes into confusable classes. FSCE freezes the backbone while keeping the FPN, RPN, and RoI head trainable during the fine-tuning over base and novel classes.

DeFRCN [8] is another method that applies the two-stage fine-tuning approach, freezing only the RCNN head. DeFRCNN focuses on solving the inconsistency between the optimization goals of class-agnostic RPN and class-aware RoI head, by introducing decoupling gradients and PCB block.

1.1.2 One-shot object detection (OSOD)

OSOD is a specific subtask of the FSOD parading. Most one-shot object detectors utilize a matching module, aiming to detect the most similar proposal to the query in the target image by measuring the similarity between them. CoAE [10] applies the non-local operation to both query and target feature maps to yield more information from the query, leading to better region proposals. They then utilize the squeeze-co-excitation block to build a relation between the feature maps of the target and the query.

SiamMask [1] leverages a matching algorithm to evaluate the similarity between feature maps of the query and the target image, which are obtained by the famous Mask R-CNN object detector [18]. They obtain the similarity feature map between the query and the entire target image. Our baseline model, BHRL [13] improves upon SiamMask [1], correlating the query and target on the instance level instead of relating the query and the entire target image. BHRL [13] utilizes a three-head relation module IHR to build robust relations between each region proposal obtained by RPN [19] and the query image.

FSOD [12] proposes the attention-RPN, building a relationship between the query and target with the depthwise convolution to generate region proposals in better quality. FSOD [12] also introduces the three-head relation head between the target and query as BHRL [13], however, it separately utilizes three relation heads, achieving three

different classification scores by each of them. On the other hand, BHRL [13] proposes a fusion layer to combine all the relation features obtained by relation heads before feeding the fused feature map to the RCNN head [19].

1.1.3 Cross-domain adaptation

Cross-domain one-shot object learning (CD-OSL) aims to enhance the generalization capacity of OSL models across diverse datasets, alleviating the challenges raised by domain shift. Although most CD-OSL studies in the literature are related to classification tasks rather than detection, recent developments have introduced benchmarks and methods targeting cross-domain FSOD (CD-FSOD). A recent study [20] establishes a comprehensive evaluation benchmark to evaluate the OSOD and FSOD models' adaptation capability to diverse datasets. For this evaluation, they prepare several datasets by focusing on increasing the diversity between them. This benchmark contributes to the field by highlighting the limitations of several methods during the cross-domain evaluations. However, despite their differences in application areas or the degree of perspective distortion, all the benchmark datasets are still based on the still-image domain.

MoF-SOD [21] is another benchmark study that aims to evaluate the OSOD and FSOD models' effectiveness across diverse datasets. They also evaluate transfer learning and meta-learning-based methods across different datasets such as medical, satellite, and natural images. Although this study highlights the performance degradation across different datasets and implicates the need for domain-agnostic approaches in FSOD and OSOD, it also uses still-image datasets for both training and evaluation, which may not provide comprehensive insights for cross-domain evaluations that focus on still-image and video domains.

On the other hand, CD-ViTO [9], proposes a novel CD-FSOD model enhancing the recent transformer-based FSOD model De-ViT [22] as well as establishing a comprehensive benchmark for future CD-FSOD studies. The study [9] establishes an evaluation benchmark comprising diverse datasets, introduces new metrics to measure the domain gap, and improves upon the De-ViT [22] across different datasets. This study distinguishes itself from the previous ones by introducing novel metrics that

measure the diversity across target domains, thereby revealing the root causes for issues caused by domain shift. Even though these new metrics are helpful, the use of still-image datasets for both training and evaluation limits the possible findings of CD-FSOD.

Distinctively, this thesis diverges from these studies by focusing on enhancing OSOD techniques in the cross-video domain using the video domain dataset as the target domain rather than the diverse still image dataset like the previous ones.

1.2 Motivation and Purpose of Thesis

OSOD models aim to detect novel classes using a single sample of a previously unseen class, enabling rapid adaptation to novel classes without extensive labeled data. However, their performances highly depend on the training domain, which is usually restricted to still images. Due to domain shift, this conventional setup yields significant performance degradation in one-shot video object detection.

Despite recent advances in assessing OSOD models across different databases, evaluations transitioning from still-image domains to video-image domains are noticeably absent in the literature. The video domain introduces specific challenges that are not found in still images, including quick alterations in the appearance of the target object, frequent occlusions, the presence of similar objects, sudden changes in lighting and background, as well as swift camera movements. These challenges result in considerable performance drops during cross-domain evaluations.

The purpose of this thesis is to first demonstrate this performance degradation and then to propose a solution to alleviate this gap effectively.

1.3 Outline

The working principles and network architecture are detailed in Chapter 2. The proposed method is discussed in Chapter 3. An overview of the datasets, an analysis of the domain gap between the training and evaluation datasets, and the performance evaluation metrics are covered in Chapter 4. Chapter 4 also delves into the experimental details and presents both visual and numerical results. Finally, we conclude the thesis in Chapter 5.

2. ONE-SHOT OBJECT DETECTION VIA MULTI-LEVEL FEATURES

In this thesis, we propose a solution for the performance gap that one-shot object detection models suffer from during cross-domain evaluations. To propose a solution, we first demonstrate the performance degradation that the Balanced and Hierarchical Relation Learning for One-shot Object Detection (BHRL) [13], a recent one-shot object detection model, faced during evaluations across different domains. In particular, the BHRL model, trained on the still-image domain MS COCO [15], is evaluated on the video dataset VOT-LT 2019 [16] to highlight the impact of domain shift on its performance. After demonstrating the performance gap caused by the domain shift, we employ the BHRL model as a baseline to develop a solution for this issue.

In this chapter, we define the problem formulation and explain the working principle of the baseline model BHRL.

2.1 Problem Formulation

The one-shot object detectors aim to adapt to the set of objects from previously unseen novel classes N with the help of a single example by transferring the knowledge gained in the learning phase performed with the set of base classes B. The base classes B included in the learning process do not overlap with those N only seen in the inference phase, where $B \cap N = \oslash$. During the inference phase, the OSOD is capable of detecting all instances of a previously unseen novel class from N within the given target image T, using only a single reference or *query shot* Q. In the scope of one-shot object detection, this reference image Q is referred to as the *query shot*.

The main challenge that one-shot object detectors suffer from is the requirement of being able to generalize from only a single novel query shot, Q that belongs to a previously unseen class, N, to detect all instances of this novel class within the target image, T. Aside from the difficulty caused by a lack of sufficient query data, there is

a notable risk of overfitting the object representation on the query shot Q in one-shot object detection. This can result in mistakes and false detections. To alleviate these problems, multi-level feature learning which enables the model to learn representations of the novel query shot Q from an unseen class N at multiple levels is proposed in the baseline model, BHRL [13]. BHRL is a recent work achieving SOTA performance in one-shot object detection by utilizing multi-level feature learning. By this approach, the model can extract a rich set of features and build a more robust and complex understanding of the novel object represented on the query shot Q. We use BHRL as a baseline study to leverage its representation capability, mainly provided by multi-level feature learning. In addition to this, we propose a novel one-shot conditional detection framework, integrating an adaptive query shot selection mechanism that is enhanced with the unsupervised online fine-tuning strategy to the BHRL.
In this section, the baseline study BHRL [13] is explained in detail.

2.2 Baseline Network Architecture

To extract visual features from both the query Q and the target T, the model combines a shared-weight Siamese ResNet-50 [23] with a Feature Pyramid Network (FPN) [24]. The model uses a typical Region Proposal Network (RPN) [19] that has been enhanced by a similarity map computed between the query and target features, which boosts the capability of generating region proposals that are more relevant to the query image Q. The features of these proposals in the target T and the extracted features of the query Q are then fed into the Instance-level Hierarchical Relation (IHR) module to achieve the multi-level semantic relation between the target T and Q. The relation map produced by IHR module is feed into the R-CNN head to detect objects.

2.2.1 ResNet-50 backbone

ResNet-50 [23] is a famous architecture that has proven to be highly effective for various computer vision tasks, including object detection. The primary innovation that ResNet-50 brings along is "residual blocks" which aims to tackle the vanishing gradient problem allowing the network to be much deeper without suffering from training difficulties. The shortcut connections found in each ResNet-50 residual block

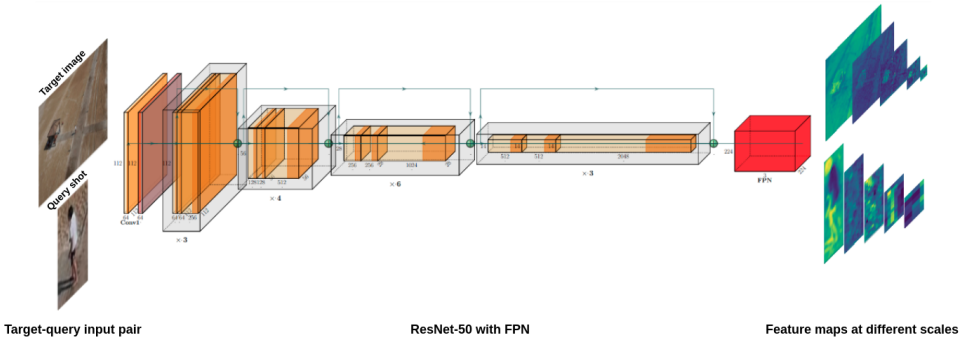

Figure 2.1 : ResNet-50 with FPN.

allow the network to learn residual functions concerning the layer inputs by skipping one or more layers. This approach allows the gradients to flow through these shortcut connections during backpropagation, minimizing the risk of gradient vanishing or exploding.

Gradient vanishing. The vanishing gradient problem is a major challenge faced during the training of deep neural networks. The issue appears if the gradients of the network's loss with respect to the parameters become very small as they are back-propagated through the network during the training phase. The learning process for the network's early layers is essentially stopped when gradients disappear since those layers' weights receive a relatively little update. This leads to an insufficient training process since the early layers, which are responsible for learning the low-level features, fail to capture informative patterns from the input data.

In the scope of object detection, ResNet-50 serves as a powerful backbone model, with an extensive capacity of producing rich feature representations for object detection. It is a commonly used feature extractor on one-shot object detection tasks, where the model needs to generalize well from a small number of samples.

2.2.2 Feature pyramid networks (FPN)

Feature Pyramid Networks (FPN) [24] is a famous module that aims to enhance the representation capability of convolutional networks in detecting objects at multiple scales. The main idea behind the FPN is to build a pyramid of feature maps at different scales, with each level of the pyramid capable of extracting high-level semantic features at its corresponding scale.

In this architecture, lower-resolution semantically stronger feature maps produced by late layers are upsampled and enriched with spatially stronger ones with higher resolution, generated by the early layers. By this approach, the spatially rich feature maps can benefit from the semantic information richness of late-layer feature maps, which improves the ability of models to detect smaller objects. This ability is particularly essential for one-shot object detection, where the model is responsible for detecting all instances of novel categories in the image using only a single sample, regardless of their scale in the image.

Combining ResNet-50's depth architecture and residual learning with FPN's multi-scale representation capabilities, these two models provide a strong architecture for the one-shot object detection task that aims to efficiently detect and classify objects from data-sparce novel classes. The combination of FPN and ResNet-50 produces multi-scale feature maps from a given input image. The baseline model BHRL [13] utilizes a shared weight siamese architecture, with two identical network branches for the target (T) and the query (Q) as seen in Figure 2.1. Note that input pairs are selected from the person14 sequence in the VOT-LT2019 video dataset [16].

2.2.3 Region proposal network

Region Proposal Network (RPN) [19] was first introduced by the famous object detection framework Faster-RCNN [19]. RPN is proposed to generate the region proposals in which a target object might be located. It operates on feature maps produced by a feature extractor, which is ResNet-50 with FPN in our case, and generates region proposals that indicate the possible regions in which the target object is located on the input feature map. RPN makes region proposals using k different regions pre-defined in terms of their aspect ratios and scales, which are called *anchors*. In particular, a window with $n \; x \; n$ is moved over the input feature map, and each of these portions is fed into the classification and regression branches after they are transformed to D dimensional vector. For each of the k anchors, the classification branch produces an objectness score indicating the probability of a target object being within the proposed region, while the regression branch produces offset values to refine these proposals. In the baseline study BHRL [13], the similarity map is computed

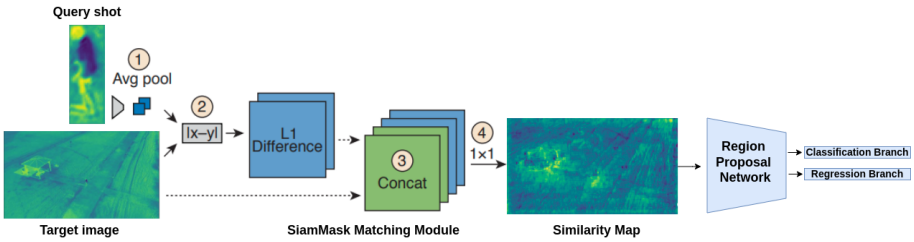

Figure 2.2 : Matching module [1] and RPN.

between the feature maps of the query Q target T, which was produced by ResNet-50 combined with FPN. The similarity map is then fed into the RPN, thereby helping the RPN to focus on the most possible regions most likely to contain the target object as seen in Figure 2.2.

2.2.4 Multi-level feature learning

The main contribution of the baseline study [13] lies in leveraging the multi-level feature learning for one-shot object detection. For this, BHRL proposes the Instance-level Hierarchical Relation (IHR) module designed to build complex relations between the query Q and the target T. The IHR module produces a relation map between the Q and the target T by fusing three different relation layers, each reflecting a different type of relation. This fused relation map, which has various levels of semantic information, is then fed into the R-CNN head. R-CNN head is utilized to detect the query-related regions using the relation map.

The IHR module has different relation layers: contrastive, salient, and attention.

2.2.4.1 Constrastive-level relation

This branch focuses on global dissimilarities between target T and query Q, emphasizing the distinctive features that distinguish the unseen novel class N present in the query image Q from other objects in the target image T. These dissimilarities are highlighted by comparing each position of the target feature map with the query feature vector.

For the comparison, the subtraction operator is used, highlighting the areas of the target image T that distinguish it from the query, indicating potential regions where the query

object might be located.

The contrastive level relation map is achieved by (2.1).

$$\mathbf{R}_c = \mathrm{P}_{\frac{C}{2}}\left(\left|R\left(AP\left(\mathbf{F}_Q\right)\right) - \mathbf{F}_T\right|\right), \tag{2.1}$$

where F_Q and F_T represent the global features of the query and the target, respectively, with each having the channel size C. $|\cdot|$ denotes the absolute value operator, whereas the $P(\cdot)$ indicates the pointwise convolution with the channel size $\frac{C}{2}$. $R(\cdot)$ is a repeat operator that makes $\mathbb{R}^{C\times 1\times 1} \rightarrow \mathbb{R}^{C\times K\times K}$, whereas the $AP(\cdot)$ denotes the average pooling operation. The equation demonstrates how the repeat operator $R(\cdot)$ aligns the dimensionalities of the query (Q) and the target (T) feature maps, making the subtraction possible.

2.2.4.2 Salient-level relation

This branch aims to identify the most unique and notable parts of instances through depthwise convolution. These salient regions are crucial to establish a relation between the query Q - target T images. This salient-level relation is established by using the query embedding as a convolutional kernel across the target feature in a depthwise manner. Depthwise convolution ensures that the generated relation feature map shares the same resolution as the target feature map, helping the model better represent saliency-based relations by keeping rich semantic information. By this approach, the target image's salient regions that most closely resemble the prominent features of the query can be identified by the model.

The salient-level relation map is achieved by (2.2)

$$\mathbf{R}_s = \mathrm{P}_{\frac{C}{2}}\left(D\left(AP\left(\mathbf{F}_q\right), \mathbf{F}_t\right)\right), \tag{2.2}$$

where $D(\cdot)$ denotes the depth-wise convolution.

Depth-wise separable convolution. In the baseline model, the similarity map between the query (Q) and target (T) is computed depth-wise manner as in previous studies [12] and [25]. Depth-wise separable convolution is a combination of the depth-wise and the point-wise convolution operations. The depth-wise convolution is responsible for identifying the spatial correlations, while the point-wise convolution focuses on capturing the cross-channel correlations between inputs. The idea behind

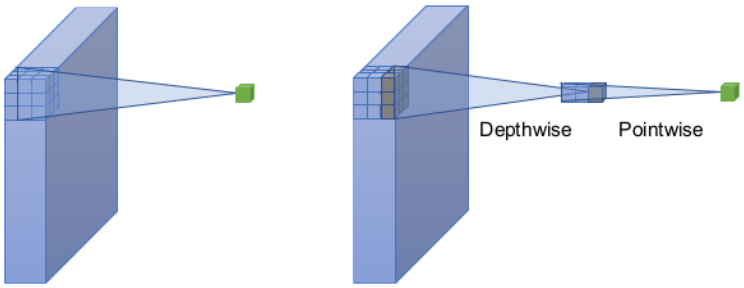

Figure 2.3 : Standard (left) and depth-wise separable convolution (right) [2].

the depth-wise separable convolution is to decouple the spatial correlations and cross-channel correlations, resulting in more useful features [26].

Depth-wise separable convolution is commonly used to reduce the computational cost and network complexity since each filter interacts with only one channel of the input. While reducing the cost and complexity, it results in each output channel carrying information solely from its corresponding input channel, limiting the ability of the model to capture complex patterns across channels. To achieve a balance between reducing the computation cost and enhancing the ability of the model to learn complexities, the point-wise convolution is applied just after the depth-wise convolution as seen in Figure 2.3. With point-wise convolution, the features from different channels can be combined, enhancing the ability of the model to learn complex patterns.

2.2.4.3 Attention-level relation

This branch focuses on capturing the local relations between the query (Q) and the target (T). The baseline model adopts the cross-attention mechanism to create spatial-aware query embedding to be compared with the target embedding at the local level.

In the implementation, local feature maps both query Q and target T are extracted. The spatial similarity is then computed between these local feature maps. To establish complex non-linear relationships between them, the softmax activation function is employed on the spatial similarity map, achieving W_S as in (2.3).

$$W_s = \text{softmax}\left(\left(\text{Conv}_{\frac{C}{8}}(\mathbf{f}_T)\right)^T \text{Conv}_{\frac{C}{8}}(\mathbf{f}_Q)\right), \tag{2.3}$$

where, $Conv_{\frac{C}{8}}(\cdot)$ denotes the convolution operation with the channel size $\frac{C}{8}$. f_Q and f_T indicate the local features of the target and the query, respectively.

W_S is applied on the global query (Q) feature as a weight to achieve a spatial-aware query feature denoted as $W_s\mathbf{F}_q$ in (2.4).

$$\mathbf{R}_A = \mathrm{Conv}_{\frac{C}{2}}\left(\left|W_s\mathbf{F}_q - \mathbf{F}_t\right|\right) \tag{2.4}$$

The local semantic relation between the spatial-aware query and the global target feature is extracted by the simple yet efficient subtraction operator.

2.2.4.4 Fusion-level relation

In the fusion level, the relation feature maps computed using the global query features, the contrastive R_C and salient-level RS relation maps, are first integrated. After concentration, the point-wise convolution is applied to reduce the dimension size to $\frac{C}{2}$. Then the attention-level relation map R_A which is produced using the local query features is concatenated. Finally, the global target feature map F_T is integrated, and another point-wise convolution with the size $\frac{3C}{2}$ is applied to achieve the desired dimension.

The fusion-level relation map $\mathbf{R}$ is achieved by (2.5).

$$\mathbf{R} = \mathrm{P}_{\frac{3C}{2}}\left(\left[\mathrm{P}_{\frac{C}{2}}\left([\mathbf{R}_s, \mathbf{R}_c]\right), \mathbf{R}_a, \mathbf{F}_T\right]\right) \tag{2.5}$$

2.2.5 R-CNN head

The R-CNN head is responsible for making final decisions based on the proposals produced by the RPN. In BHRL [13], the relation map produced by the IHR module is fed into the R-CNN head to detect the region proposal which is the most similar to the query shot Q. This is possible because the IHR builds a relation map that encapsulates the complex relationships between the region proposals within the target T and the query shot Q, thereby enabling the R-CNN head to focus on detecting the region that is most similar to the query shot Q. Feature maps of each proposal are fixed to the same size using the RoI Align layer and fed into the sequence of fully connected layers to achieve a fixed-size vector. This fixed-size vector is then

fed into two branches that work in parallel, classification, and regression branches. After predictions are finalized, the Non-Maximum Suppression (NMS) is employed to eliminate overlapping predictions, using the overlapping threshold of 0.5.

2.2.5.1 Classification branch

The classification branch is responsible for predicting the class label of the corresponding region proposal, including the background, producing a probability distribution over all the possible classes. The BHRL employs binary classification, utilizing class labels to denote the presence or absence of the target object within the proposed region. In the BHRL, predictions with a confidence score lower than the score threshold of 0.05 are ignored. For the binary classification tasks, the Binary cross-entropy (BCE) loss function is employed. It measures the difference between the ground-truth labels and the predicted probabilities of the positive class. For a single prediction, the binary cross-entropy loss is calculated as in (2.6):

$$L_{cls}(y,p) = -[y\log(p) + (1-y)\log(1-p)], \tag{2.6}$$

where y denotes the ground-truth label, whereas p is the probability of the prediction being in the positive class. In the binary classification tasks, ground truth can take only two values, 1 or 0, indicating the existence or absence of the target object, respectively.

2.2.5.2 Regression branch

The regression branch is responsible for predicting the object location in the corresponding region proposal. The regression branch produces four numbers, which are used to refine the region proposals. To refine the region proposals, Smooth L1 loss is used as in (2.7).

$$L_{\text{reg}}\left(z,\hat{z}\right) = \sum_{i\in\{x,y,w,h\}} \text{L1}_{smooth}\left(z_i - \hat{z}_i\right), \tag{2.7}$$

where, z and $\hat{z}_i$ denote the predicted BBOX coordinates and the ground-truth coordinates. Smooth L1 loss is given in 2.8.

$$\text{L1}_{smooth}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{2.8}$$

The final loss for the training is a weighted combination of the classification and regression branch losses as in (2.9).

$$L_{total} = \lambda L_{cls}(y, p) + L_{\text{reg}}\left(z, \hat{z}\right) \tag{2.9}$$

The λ is taken as 1 in the baseline study.

2.2.6 Dataset settings

The baseline model uses MS COCO [15] and Pascal VOC20017-2012 [17] datasets for both the training and evaluation phases. The study divides 80 classes from the COCO dataset into four equal parts. They take one split as novel N, whereas the remaining three of them as base classes B each time during the experiments. For the Pascal VOC, they split the dataset into two parts, 16 classes as base classes B, and 4 classes as novel classes N. In the evaluation phase, they investigate the AP_{50} performance of the model on both base B and N classes.

2.2.6.1 Target-query matching

During the training phase, a target image T with a seen class instance B is randomly chosen, as is a query image Q with that class. For the evaluation, five different query images Q for each class from $B \cup N$ that is present in the target image T are chosen to be tested five times, and average AP_{50} scores are reported.

The baseline model [13] does not conduct any cross-database evaluation and uses PASCAL [17] and COCO [15] datasets for both training and evaluation as in the majority of one-shot object detectors [10], [1], [11]. This common use in the literature arises from the limitations set by the nature of the one-shot object detection. Most one-shot object detectors utilize a template matching algorithm between the query Q and the target T. By this approach, they achieve satisfactory results in the still image domain where images are stationary and identifiable while being vulnerable to the challenges caused by the domain shift [14]. This vulnerability encourages the research community to use databases sharing similar data distributions and class characteristics for both training and evaluation. We mitigate this vulnerability by proposing to integrate an unsupervised online fine-tuning mechanism to the inference architecture of the baseline model BHRL [13].

3. CROSS-DOMAIN ADAPTATION BY TRANSFER LEARNING

The main motivation for this thesis is to address and propose a solution for the performance degradation that one-shot object detectors face during the cross-domain evaluations caused by the challenges that domain shift brings along. After analyzing the reasons for this gap, we propose a novel "One-shot Cross-domain Adaptation by Online Fine-tuning" framework (OSCDA). This framework improves upon the BHRL [13], which is the baseline model for this thesis, by incorporating an unsupervised online fine-tuning approach, complemented by an adaptive query shot selection mechanism.

This section provides a detailed explanation of the proposed methods.

3.1 One-shot Cross-domain Adaptation by Online Fine-Tuning (OSCDA)

Recent transfer learning-based methods offer promising solutions for transferring the data-abundant base (B) class knowledge to data-scarce novel classes (N), utilizing several methods such as increasing the number of positive novel samples [6], introducing contrastive loss [7], and adding the gradient decoupled layer and an extra calibration block to achieve a balance between the regression and classification tasks and boost the classification accuracy, respectively [8]. However, these methods become inefficient in transferring the base class knowledge to a novel class coming

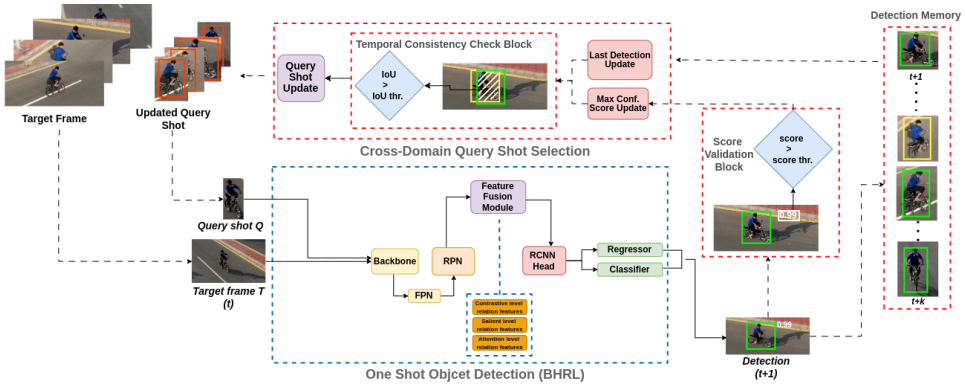

Figure 3.1 : One-shot cross-domain adaptation by online fine-tuning (OSCDA).

from a cross-domain as they cannot overcome the challenges raised by the domain shift [14].

On the other hand, the baseline OSOD method BHRL [13] claims that it does not need any fine-tuning process, to achieve satisfactory performance specifically on the same evaluation domain as the one used for training. Yet, BHRL lacks cross-domain performance evaluation. In the scope of this thesis, first, we demonstrate that it is vulnerable to severe domain shifts by evaluating BHRL on the cross-domain video dataset VOT-LT 2019 [16].

We alleviate this vulnerability by jointly fine-tuning the BHRL network [13] trained on the Split3 base classes from still-image dataset COCO [15], on a single query shot (Q) from the novel class N coming from the cross video domain VOT-LT 2019 [16], in the inference phase as seen in Figure 3.4. We introduce the Cross-domain Query Shot Selection (CDQSS) layer incorporated with the baseline method [13], for the adaptive query shot (Q) selection. When the CDQSS is enabled, a new query shot Q is adaptively selected based on the model's prior decisions during the inference. Subsequently, the model undergoes online fine-tuning using the chosen query shot Q taken from the cross-domain dataset VOT-LT 2019 [16] during inference. This fine-tuning process is unsupervised, as it relies solely on the model's prior decisions rather than ground truth data.

The new query shot Q is adaptively chosen based on specific predefined conditions within a K-frame range. To evaluate these conditions, the confidence scores of model predictions are continuously monitored across K-frame segments throughout the video sequence. After the model's predictions are validated based on the confidence score, the temporal consistency between the prediction with the highest score and the most recent one is checked by the Temporal Consistency Check Block using the Intersection over Union (IoU). If all the conditions are satisfied, the chosen query shot Q is used to adapt the model to the updated appearance represented by itself, by applying unsupervised fine-tuning. In the evaluations discussed in Chapter 4, we examined the model's performance both with and without the activation of the CDQSS, referred to as OSCDA w/ CDQSS and OSCADA w/o CDQSS, respectively.

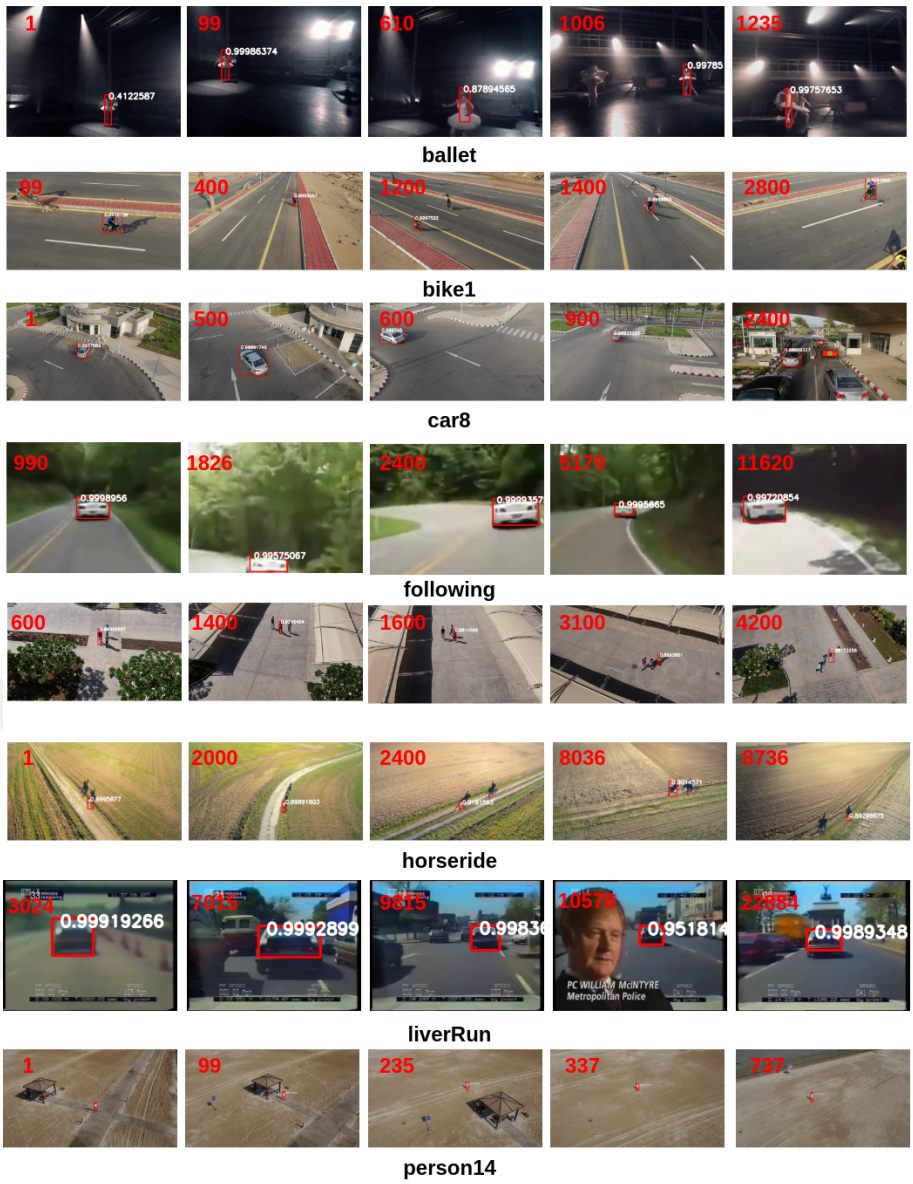

Figure 3.2 : Query shot updates by CDQSS.

For visual inspection, a number of the query shots adaptively chosen during the inference are given with the corresponding frame numbers in Figure 3.2. As seen in 3.2, by continuously selecting new query shots during the inference, CDQSS allows the model to capture the target object across various aspects and scales. Notably, the CDQSS demonstrates the ability to consistently choose the target object as query shot Q in long-term video sequences, such as liverRun or horseride, despite significant changes in the appearance of the target object.

3.1.1 Loss function

We keep all the network parameters trainable during the cross-domain fine-tuning. During the fine-tuning, we employ the joint loss function for regression and classification tasks given in Eq. (3.1) as in the baseline study BHRL [13],

$$L = L_{L1} + L_{BCE} \tag{3.1}$$

where L_{L1} represents a smoothed L1 loss used in the regression branch, and L_{BCE} denotes the binary cross-entropy loss (BCE) applied in the classification branch, as given in Figure 3.4.

The smoothed loss given in Eq. (3.2) aims to minimize the mean absolute difference between BBOX coordinates (y_i) of the novel query shot and the BBOX coordinates ($\hat{y}_i$) predicted by the model where N_s denotes the number of shots.

$$L_{L1} = \frac{1}{N_s} \sum_{i=1}^{N_s} |y_i - \hat{y}_i| \tag{3.2}$$

For the classification branch, L_{BCE} is computed as in Eq. (3.3),

$$L_{BCE} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \left[p_{y_i} \log(p_{\hat{y}_i}) + (1 - p_{y_i}) \log(1 - p_{\hat{y}_i}) \right] \tag{3.3}$$

where p_{y_i} and $p_{\hat{y}_i}$ indicate the actual class label (either 0 or 1) of the target y_i and predicted probability of positive class (ranging from 0 to 1), respectively.

The one significant reason why we keep all the network parameters trainable during the fine-tuning is to leverage the advantages of multi-level feature learning by allowing them to continue to learn from the novel object representation. We initialize the weights from [13], which were pre-trained on the COCO dataset [15] and proceed to fine-tune the model with a single example taken from the challenging video domain VOT-LT 2019 [16]. Unlike most one-shot object detectors that use both base and novel classes during the fine-tuning, we use only novel classes for the fine-tuning. The fine-tuning process continues throughout the video sequence using different one-shot examples adaptively chosen from the model's detections as in the Cross-domain Query Shot Selection (CDQSS) block in Figure3.4.

3.1.2 Cross-domain query shot selection (CDQSS)

Initially, the novel query shot Q is specified as the ground truth representation of the target object in the first frame of the video sequence (y_1) and the model is subsequently fine-tuned using this chosen shot. During the initial fine-tuning, the number of samples is set to 1, and the L1 loss is used to minimize the absolute difference between the predicted ($\hat{y_1}$) and actual bounding box coordinates in the first frame (y_1), as in Eq. (3.4).

$$L_{L1} = |y_1 - \hat{y}_1| \tag{3.4}$$

BCE loss is also used to minimize the difference between the predicted probabilities and the actual class label ($p_{y_1} = 1$) of the target in the first frame across epochs, as in Eq. (3.5).

$$L_{BCE} = -\log(p_{\hat{y}_1}) \tag{3.5}$$

However, integrating online fine-tuning with an OSOD brings along the risk of overfitting considering that only a single query shot is used for fine-tuning. Moreover, keeping all the parameters trainable during the fine-tuning also increases the possibility for the model to overfit on the novel query shot Q. To mitigate this risk, we adaptively update the query shot across the video sequence with the model's detections using the Cross-domain Query Shot Selection method.

We store the model's decisions in the detection memory across K-frame segments throughout the video sequence to monitor their confidence scores. K is an adjustable parameter, we choose a safe threshold by setting K as 100. The Score Validation Block then validates each detection, filtering the one with a confidence score lower than the score threshold which is set as 0.85. The parameter set at 0.85 is adjustable. We chose this value because the model usually yields high confidence scores, leading us to use a high score threshold to confirm that the model's predictions are true positives. By this approach, we prioritize the predicted confidence scores achieved by the baseline model [13], leveraging the advantages of the multi-level feature learning approach.

Most challenges of the video domain adaptation are raised by the rapid appearance changes of both target and background occurring across the video sequence. Even

if the model is trained to detect the most similar proposal to the query shot Q in the target image T, the appearance of the query might significantly change throughout the video sequence, leading the model to make poor predictions. We alleviate this problem by introducing the Temporal Consistency Check Block that evaluates the temporal alignment between the latest detection and the most confident one. If the Intersection over Union (IoU) between them is higher than the threshold which is pre-defined as 0.70, the query shot Q is updated to the most confident detection. In this scenario, we apply unsupervised fine-tuning, where the model's detection (y_k^{best}) selected by the CDQSS from the frame $k< K$ is set as ground truth. The L1 loss is employed to minimize the difference between the selected detection and the predicted detection ($\hat{y}_k$) at frame k across the epochs as in Eq. (3.6).

$$L_{L1} = \left| y_k^{best} - \hat{y}_k \right| \tag{3.6}$$

BCE loss is also used to minimize the difference between the predicted class probabilities and the actual class label of the model's detection selected by the CDQSS ($p_{y_k^{best}} = 1$) across epochs, as in Eq. (3.7).

$$L_{BCE} = -\log(p_{\hat{y}_k}) \tag{3.7}$$

If the pre-defined conditions are not met, the query shot is reverted to the initial query shot. While this method doesn't entirely mitigate the challenges raised by rapid appearance changes, our experiments show that relying on the initial provided ground truth representation of the target object in these cases significantly minimizes the issues associated with the appearance shifts in the temporal domain.

We evaluate the proposed method with and without the CDQSS to see the contribution of updating the query shots across video frames. Without CDQSS, the model is fine-tuned once with the specified number initial query shot only at the beginning of the sequence .

K is an adjustable parameter that allows us to update the query shot Q on which the model is fine-tuned from the model's decisions. We ensure that the chosen query shot Q is a true positive by evaluating it based on different conditions such as their confidence scores and temporal consistencies. This strategy helps prevent overfitting to the initial query shot derived from the ground truth. Obviously, if the model's decision chosen to

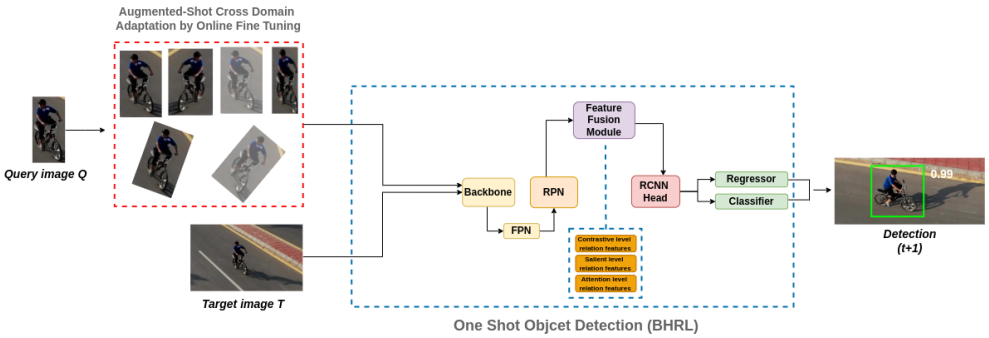

Figure 3.3 : Shot augmentation for cross-domain adaptation (SACDA).

be updated as new query shot Q is not a true positive, the model would fail to detect the instances of it on subsequent frames. For this very reason, we maintain a safe score threshold and propose the framework OSCDA (w/ CDQSS) such that it updates the query shot, frequently.

3.2 Shot Augmentation for Cross-domain Adaptation (SACDA)

Data augmentation is a widely known method to increase the diversity of the data to be consumed by deep learning models, making the models more robust to appearance changes. It also serves to minimize the risk of overfitting since acting like a regularizer during the training.

In the proposed method OSCDA (w/o CDQSS), fine-tuning the model only with the initial appearance of the target carries a high risk of overfitting. On the other hand, the continuous end-to-end fine-tuning on different shots across a sequence brings along a time consumption even though we use a simple yet efficient loss function for optimization during fine-tuning. To achieve a balance between reducing time consumption and minimizing overfitting risks, we introduce the Shot Augmentation for Cross-domain Adaptation (SACDA) approach that enhances the initial query frame's diversity by applying widely used data augmentation techniques in conventional object detection as seen in Figure 3.3. In this way, we allow the model to adapt to these synthetically generated shots during fine-tuning, which is performed only once at the beginning of the inference phase. We increase the model's ability to learn different variations of the initial query shot without updating the query shot at various times with different appearances and scales. We do not incorporate any query shot update and continuous fine-tuning process, fine-tuning the model only at once at the beginning

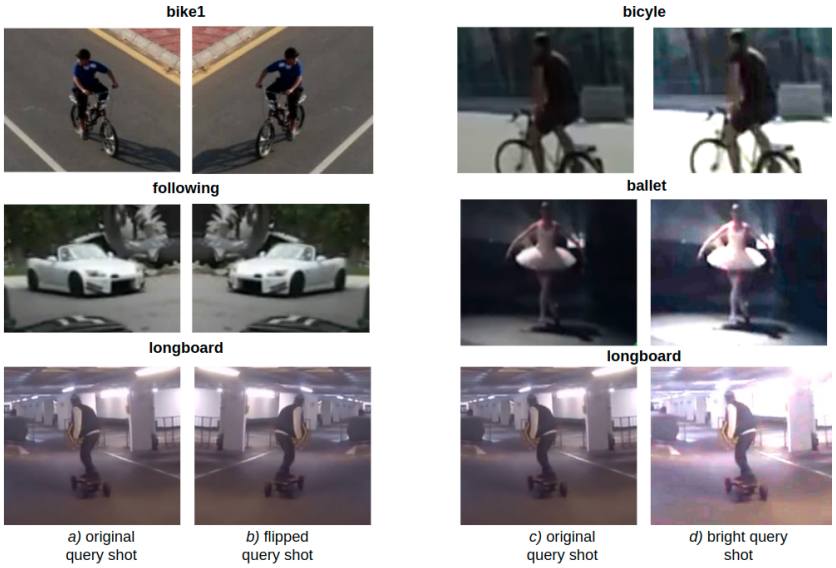

Figure 3.4 : Augmented query shots.

of the inference with the initial query shot's variations that we call "augmented shots". The main purpose of the SACDA method is to mimic the motion variations of the target as well as the illumination changes, which naturally occur through the sequence considering the complexity of the challenging VOT-LT2019 dataset [16]. We apply horizontal flip, random scale, random translation, and shearing to replicate the appearance changes of the target. Additionally, brightness adjustments are employed to replicate the scene's illumination variations during the sequence. We achieve 8 "augmented shots" ($p_{y_i} = 1$) as the result of the SACDA method.

Hence the model is only fine-tuned at the beginning of the inference phase with the augmented shots employing the L1 and BCE losses as in Eq. (3.8) and Eq. (3.9), respectively.

$$L_{L1} = \frac{1}{8}\sum_{i=1}^{8} |y_i - \hat{y}_i| \tag{3.8}$$

$$L_{BCE} = -\frac{1}{8}\sum_{i=1}^{8} \left[\log(p_{\hat{y}_i})\right] \tag{3.9}$$

We compare the Shot Augmentation for the Cross-domain Adaptation method with the One-shot Cross-domain Adaptation by Online Fine-Tuning to evaluate whether utilizing different variations of the initial query shot could eliminate the time-intensive process of continual target updates without compromising performance.

Table 3.1 : Comparison of inference time between the proposed methods.

Method	Inference time (mins)
Baseline [13]	~ 4.5
OSCDA (w/o CDQSS)	~ 8.5
OSCDA (w CDQSS)	~ 120
SACDA	~ 14.5

Discussion. We introduce the SACDA method to introduce a candidate framework that will be achieve at least the same performance as the OSCDA (w/ CDQSS) method while eliminating the time consumption that it brings along. However, as discussed in Section 4, the SACDA method yields lower AP0.5 results compared to the OSCDA (w/ CDQSS) method, with reductions of approximately 4% and 14% for base and novel classes, respectively, when tested on the COCO [15] and Pascal VOC [17] datasets. Compared to the OSCDA (w/o CDQSS), it produces approximately 1% lower results on base classes, with both models initiated with COCO and Pascal VOC datasets. For novel classes, the SACDA method shows a decrease in AP0.5 results by 3% and 8% for models initiated with the COCO and Pascal VOC datasets, respectively.To sum up, the SACDA method could not achieve either the OSCDA (w/o CDQSS) or OSCDA (w/ CDQSS) methods in any configuration. It just outperforms the other methods on specific sequences as discussed in Section 4.

Table 3.1 shows the inference time of the proposed methods on a video sequence with an average duration ($\sim$ 3000 frames). Compared to OSCDA (w/ CDQSS), the only advantage that stands out with the SACDA method is the time efficiency it provides. However, its inference time is much longer than the OSCDA (w/o CDQSS) whereas its performance is lower than it.

Considering the overall results, the SACDA method is not the preferable one based on either performance or time efficiency. As explained in Section 4, it outperforms the other methods on specific video sequences which is promising for future studies. With targeted augmentation methods, it is possible to improve the AP0.5 results without significantly increasing the inference time.

4. PERFORMANCE EVALUATION

This chapter provides information about the dataset used in the training and evaluation (Section 4.1) phases as well as the performance evaluation metrics (Section 4.3). Besides, Section 4.2 illustrates the domain gap that OSOD models experience during the cross-domain evaluations.

4.1 Datasets

This section provides a detailed examination of the datasets used for both the training and the cross-domain evaluation phases, along with an analysis of the domain gap between them.

4.1.1 COCO dataset

The COCO (Common Objects in Context) dataset [15] is one of the most popular datasets for the object detection task in the community. It has been commonly used for both the training and evaluation phases of deep learning models. COCO provides a rich set of annotations for both segmentation and detection tasks. In the scope of this thesis, we are only interested in object detection annotations. The dataset contains 80 object categories with approximately 200K annotated, covering a wide range of objects from everyday life. The dataset contains 80 distinct object categories which include a broad spectrum of objects commonly encountered in daily life.

In this thesis, the pre-trained model on the particular split of the COCO dataset, split 3 as given in [1], is used for the experiments. The classes that are not included in this split, which means that the novel (N) classes are given in Table 4.1. We intentionally chose the model pre-trained on split 3 because we aimed to select a split that includes classes overlapping with those in the VOT-LT 2019 dataset, which we employed for cross-domain evaluation. This approach allows us to evaluate the model's performance

Table 4.1 : COCO split 3 novel classes.

Class id	Class
3	Car
7	Train
11	Fire Hydrant
15	Bird
19	Sheep
23	Zebra
27	Handbag
31	Skis
35	Baseball bat
39	Tennis racket
43	Fork
47	Banana
51	Broccoli
55	Donut
59	Potted plant
63	TV
67	Keyboard
71	Toaster
75	Clock
79	Hair drier

on both novel and base classes. By doing so, we aim to ensure that not all classes from the VOT-LT 2019 dataset are treated as novel in our evaluations.

The examples for the COCO dataset are given in Figure 4.1.

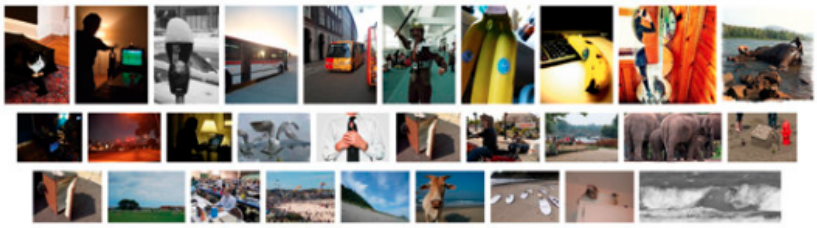

Figure 4.1 : Examples from the COCO dataset [3].

4.1.2 Pascal VOC dataset

The PASCAL Visual Object Classes [17] is a popular benchmark for object detection, providing a representative annotated dataset for object detection studies. In the baseline study, two versions of the dataset are used, VOC2007 and VOC21012. VOC2007 and V0C2012 have approximately 5k and 12k images, with 25k and 27k annotated objects, respectively.

The dataset contains 20 different object categories as seen in Figure 4.2.

Figure 4.2 : Examples from the Pascal VOC dataset [4].

In this thesis, the model pre-trained on the specific split of the Pascal VOC dataset, as given in [1], which contains 16 of 20 classes. The remaining 4 classes which are cow, sheep, cat, and aero are not included in the base training of the BHRL model.

4.1.3 VOT-LT 2019 dataset

The VOT-LT 2019 (Visual Object Tracking Long Term) Challenge is a well-known benchmark for the evaluation of visual object tracking algorithms. The challenge provides 50 different long-term video sequences, with 50 diverse scenarios containing several difficulties. The main purpose of the challenge is to evaluate the performance of tracking algorithms on these sequences where the target object may disappear for a while. Besides the target disappearances for a while, the video sequences have several difficulties like rapid appearance changes of the target object, swift shifts in the background illumination, and rapid camera movements resulting in blurry view.
The examples from 50 video sequences with their names are given in Figure 4.3. Since the benchmark was performed to evaluate the trackers, not the detectors, the annotations contain BBOXes of the target object in each video sequence but do not specify the class of the target objects. In this thesis, we use the VOT-LT 2019 dataset for a cross-domain evaluation, considering each sequence has a different class of object. Since the VOT-LT 2019 dataset serves a different purpose compared to the datasets used in the training phase Pascal VOC and COCO datasets, its distribution significantly varies from those of the training datasets. This domain shift leads to a notable degradation in the performance of the baseline model BHRL during cross-domain evaluation. Before demonstrating this performance gap and proposing a

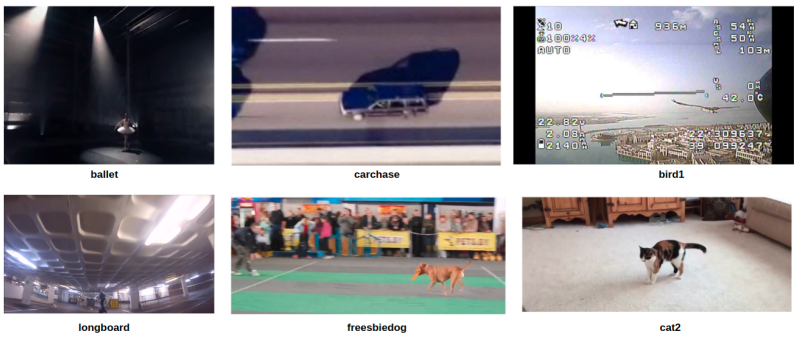

Figure 4.3 : Examples from the VOT-LT 2019 dataset [4].

solution for it, we first examined the domain gap between the training and evaluation domains in Section 4.2.

4.2 Analysis of Domain Gap

The main motivation of this thesis is to propose a solution to alleviate the performance degradation raised by the domain gap. For this purpose, we first examine this performance gap between the training and evaluation domains in this section.

We use the Inter Class Variance (ICV) metric proposed in a CD-FSOD [9] which is a very recent work studied on the cross-domain evaluation of few-shot object detection (FSOD) models. In this work, the authors demonstrate the performance degradation of the recent transformer-based FSOD model De-Vit [22] on the cross-domain datasets. Unlike us, they use the still-image domain for both training and evaluation, yet they select target datasets, for the cross-domain evaluations, with distributions that significantly diverge from the training dataset. To demonstrate the domain gap between the training and evaluation (target) domains, they propose metric ICV which defines the divergence within the classes of the respective domain. This means that the datasets in which the classes are very distinct from each other have higher ICV values and lower values when the opposite is true.

They first calculated the score matrix S for each dataset using the text features obtained by the CLIP [27] as in (4.1).

$$S = \mathscr{F} \cdot \mathscr{F}^T, \tag{4.1}$$

where $\mathscr{F}$ denotes the concatenated matrix of text features $(f_1, f_2, \ldots, f_N)$ corresponding to each class $(\{c_1, c_2, \ldots, c_N\})$ within the dataset that comprises N distinct classes.

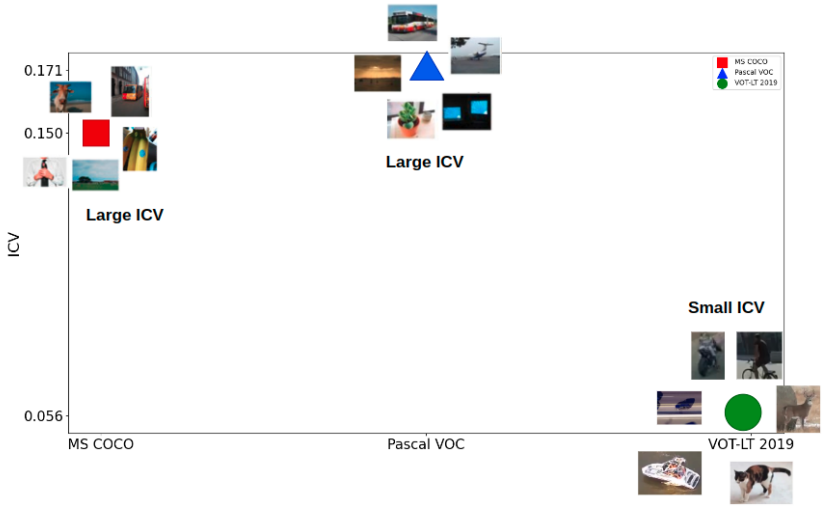

Figure 4.4 : ICV values of datasets.

Here c values indicate the name of each class.

Then they achieve ICV value for each dataset using the score matrix as in (4.2).

$$ICV = \frac{1}{N^2 * D} \sum_i \sum_j S_{ij}, \tag{4.2}$$

where D denotes the dimension size of the text features.

We follow the same steps to calculate the ICV values for training datasets, COCO [15] and Pascal VOC [17], and evaluation dataset VOT-LT 2019 [16]. As previously mentioned, the VOT-LT 2019 dataset provides only the bounding box (BBOX) coordinates for each target object in the corresponding sequence without class labels. To address the absence of class annotations in the VOT-LT 2019, we utilized ChatGPT-4 to create class descriptions for each target object. [28].

We perform the same calculations as [9], resulting ICV values for each dataset as in Figure 4.4. Note that, for each training dataset, we only use the classes that are included in the training phase.

Figure 4.4 clearly illustrates a significant gap in the ICV values of the datasets used in the training and evaluation phases. As stated in [9], ICV measures the similarity between the semantic labels of the objects within the corresponding dataset. Such a difference between the ICV values is expected, given that the training datasets consist of still images, while the evaluation dataset originates from the video domain, highlighting the divergence of the dataset characteristics. The authors [9] state that

transferring knowledge between datasets with high ICV values, like the datasets used for the training of the BHRL model, is relatively straightforward. However, transferring knowledge to a cross-domain dataset with a low ICV, where the distinction between classes is less pronounced as in the VOT-LT 2019 dataset used for the cross-domain evaluation, presents a significant challenge. In light of this information, this variance between the ICV values of the training and evaluation datasets provides a rational basis for the observed performance degradation of the BHRL model in cross-domain validation scenarios.

4.3 Evaluation Metrics

Mean Average Precision (mAP) which is a commonly used metric for the performance analysis of one-shot object detection models in literature is used to evaluate the detection performance of the model. In particular, we use the mAP calculated at an IoU threshold of 0.5, which is referred to as mAP0.5.

4.3.1 Intersection over union (IoU)

Intersection over Union (IOU) measures the overlapping area between the model prediction and the ground truth as in (4.3).

$$IoU = \frac{\text{area}(\hat{T} \cap T)}{\text{area}(\hat{T} \cup T)}, \tag{4.3}$$

where $\hat{T}$ and T denote the predicted and the ground-truth BBOXes, respectively.
A predicted BBOX is considered a match with a ground truth BBOX if their Intersection over Union (IoU) exceeds a predetermined IoU threshold. Although a predicted BBOX can only be matched with one ground truth, a single ground truth can match with several model predictions.

4.3.2 Precision

The prediction with the highest confidence score among all predictions matched with the ground truth is considered a true positive (TP), meaning that the model has successfully identified an actual object. The other predictions that are matched with

the same ground truth BBOX are considered false positives (FP), indicating that the model made incorrect BBOX predictions for objects that do not exist in real.

The percentage of the correct model predictions is measured by the precision metric (P), as given in (4.5).

$$P = \frac{TP}{TP + FP} \tag{4.4}$$

4.3.3 Recall

The ground truth BBOXes that do not have any assigned predictions are classified as false negatives (FN), suggesting that the detector has failed to identify the existence of a real object.

Recall metric (R) measures the proportion of existing objects that are correctly detected by the model.

$$R = \frac{TP}{TP + FN} \tag{4.5}$$

4.3.4 mAP0.5

Average Precision (AP) is the area under the precision-recall curve at the specific IoU threshold. In the experiments, we use the IoU threshold of 0.5.

AP is calculated as in (4.6)

$$AP = \int_0^1 p(r)\, dr, \tag{4.6}$$

where $p(r)$ is the measured precision at the recall level r.

The mean average precision (mAP) is used to evaluate the overall detection performance of the model across all classes. It represents the average of the computed Average Precision (AP) values for each particular class.

4.4 Experiments and Results

This section provides the evaluation strategy and experimental settings used to evaluate the proposed methods in this thesis. We first demonstrate the performance degradation that one-shot object detection methods face during the cross-domain evaluation in Section 4.4.1. We then proceed to evaluate the performance of our proposed one-shot object detection framework, OSCDA, presenting in-depth experimental findings in Section 4.4.3. Finally, we extend our evaluation by including detailed visual results

and comprehensive analysis in Section 4.5. During these experiments, we use the object detection evaluation metrics explained in Chapter 4.

4.4.1 Demonstration of performance gap during cross-domain evaluation

The performance of one-shot object detection models highly depends on the training data domain mainly composed of still images from publicly available object detection datasets. As a result, a significant performance degradation raised by the domain shift is observed during the cross-domain evaluation.

We demonstrate this gap by performing cross-domain evaluations on the provided BHRL models [13] pre-trained on still-image sets, Pascal VOC [17] and COCO [15] (split 3) datasets, on the cross video dataset VOT-LT 2019 [16]. Since the OSOD models have been evaluated on the public object detection datasets, the frames in which the target object does not exist are removed from the video sequences of the VOT-LT 2019 dataset, for a fair comparison.

Class-based reporting. In the video domain, the main purpose is to successfully detect the target object, which is introduced at the beginning of the sequence, in all subsequent frames. This goal is integral to our study and is incorporated into the one-shot object detection task formulation as outlined in Section 2.1. By adopting this methodology, we treat each target object in each of the 50 video sequences from the VOT-LT 2019 dataset as a separate class. Our framework, OSCDA, is designed to detect all instances of a specific target object in every frame throughout the respective video sequence. For a fair comparison with the baseline study, we categorized the target objects in each video sequence into the base (B) and novel (N) classes. To achieve this, we first grouped each target object under the main classes to align them with those found in the COCO and Pascal VOC datasets. For instance, the target object "person" appears in 24 different video sequences, each representing varied behaviors and unique appearances. In one sequence, the person might be playing basketball, while in another, she/he could be skydiving in a winged suit. We grouped all these instances under the "person" category to ensure consistency and fairness in the performance comparison with the baseline study.

For the evaluation on the still-image domain, COCO, and Pascal VOC datasets, we

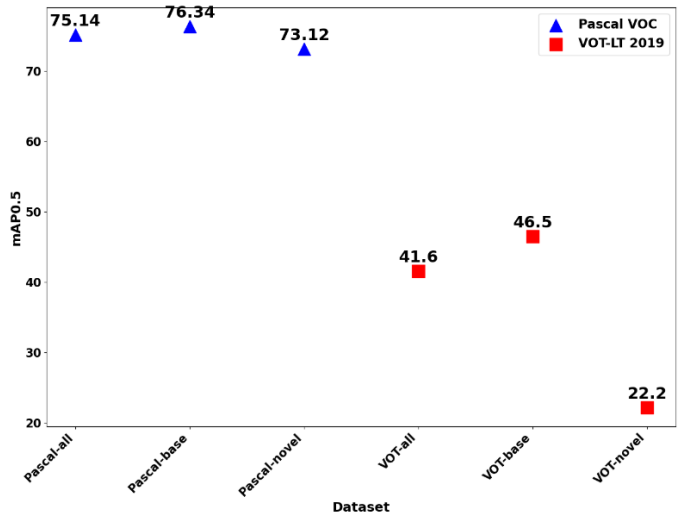

Figure 4.5 : Cross-domain evaluation of VOC model.

use the same experiment settings as the baseline model BHRL as explained in Section 2.2. During evaluation in the cross-video domain, we randomly select 5 query shots (Q) from the corresponding video sequence to test the baseline model specifically on this sequence, treating each target object in each sequence as a distinct class. We then calculate the average performance across 5 different runs for each sequence to report the mAP0.5 performance individually. Finally, we calculate the average mAP0.5 scores of sequences belonging to the base (B) and novel (N) classes separately, according to the target objects' main classes.

VOC model evaluation. We first evaluate the BHRL model pre-trained on Pascal VOC, which is referred to as the VOC model in further analysis, on test data from the cross-video domain. Normally, the VOC model [13] uses 16 classes as the base and assigns the remaining 4 classes as novel. For a fair comparison, we report the mAP0.5 performance exclusively for the base and novel classes that are shared between the Pascal VOC and VOT-LT 2019 datasets.

Figure 4.5 compares the VOC model's performance on its training domain (indicated by the blue triangles) and cross-video domain VOT-LT 2019 (represented by the red squares).

As seen in Figure 4.5, the VOC model achieves a quite sufficient object detection performance, 75.14% across all classes, on the same evaluation domain as its training. However, significant performance degradation is observed when this model

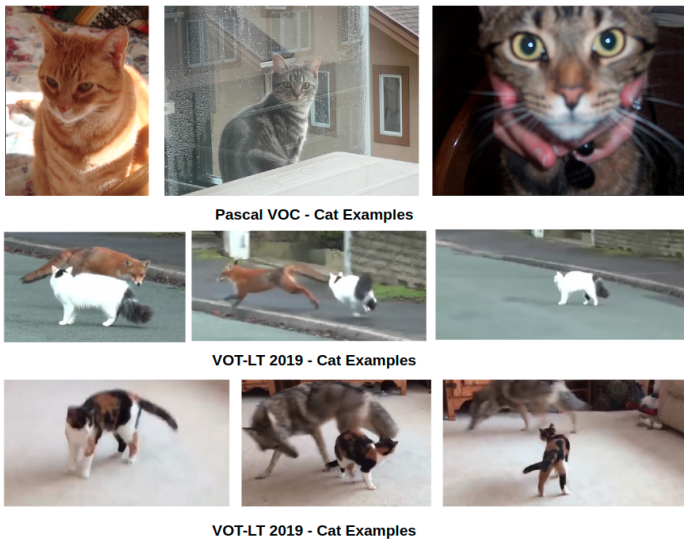

Figure 4.6 : Difference in novel classes across training and evaluation domains.

is evaluated in the cross-domain video dataset VOT-LT 2019, dropping to 41.14% across all classes. Besides, the performance gap raised by the domain shift reaches approximately 51% across novel classes, emphasizing the main motivation behind this thesis. To further investigate this performance gap, we examined the differences in data distribution of images from the same novel class, specifically cats, as shown in Figure 4.6.

As demonstrated in Figure 4.6, there is a noticeable variance in object appearances within the same class. This divergence is particularly pronounced in the cross-video domain, which presents unique challenges such as occlusion, blur, and rapid changes in appearance, which are the typical challenges of the video domain. These challenges are not typically encountered in still-image datasets like Pascal VOC where the images are still and clearly visible. These challenges are the key factors contributing to the significant performance drop observed during the cross-domain evaluations.

COCO model evaluation. We also demonstrate the performance gap and evaluate the BHRL model pre-trained on the COCO dataset, which is referred to as the COCO model, on test data from the cross-video domain. Originally [13], the COCO split 3 models uses 20 classes as the base and assigns the remaining 60 classes as novel. Consistent with the strategy used in the prior experiment, we only report the mAP0.5 performance on those base and novel classes that are common to both the COCO and VOT-LT 2019 datasets, ensuring that our evaluations are comparable and relevant.

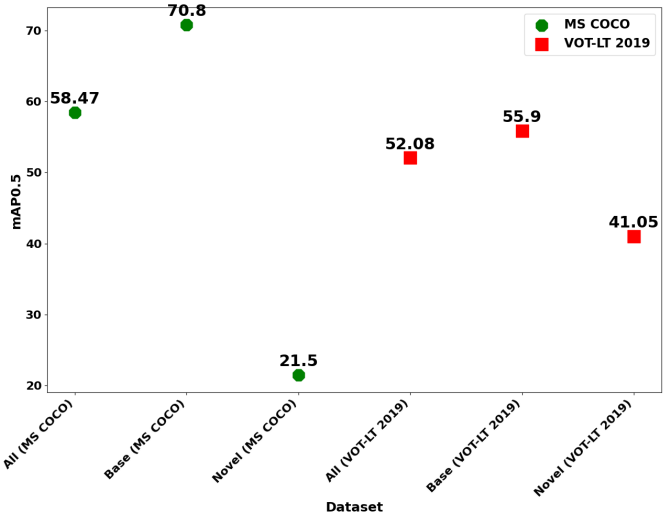

Figure 4.7 : Cross-domain evaluation of COCO model.

Figure 4.7 compares the COCO model's performance on its training domain (indicated by the green circles) and cross-video domain VOT-LT 2019 (represented by the red squares). As seen in Figure 4.7, the COCO model achieves 58% mAP0.5 performance on the same domain as its training, while it experiences a decrease to 52% mAP0.5 performance on the cross-video domain. Besides, there is a notable performance gap of nearly 15% between the image and video domain evaluations across base classes. One interesting result is that the COCO model shows better performance in the video domain across the novel classes. However, it is more reasonable to compare the mAP0.5 performance across all classes between the image and video domains, rather than solely among novel classes. This is particularly fairer because only two classes, bird and car, are common between the COCO and VOT-LT 2019 datasets.

4.4.2 Cross-domain evaluation of proposed methods

In this section, we evaluate the proposed framework OSCDA and our other experimental approach SACDA on the cross-video domain VOT-LT 2019 dataset [16]. We evaluate the proposed framework on the video dataset VOT-LT 2019 by initiating two provided models [13] pre-trained on the still-image datasets COCO and Pascal VOC, which are referred to as the COCO and VOC models. The cross-domain evaluations are performed separately for the base (B) and novel (N) classes, and there are variations in the base and novel classes between the experiments performed with

the COCO and VOC models. The categorization of novel and base classes for each model is determined based on the settings of the training dataset used in each respective experiment. As discussed in Section 4.4.1, we treat each target object in each video sequence as a distinct class, categorizing them into base and novel classes based on the main classes they are grouped under. In this section, we report the mAP0.5 performance for each sequence individually and also provide the average mAP0.5 results for both the base (B) and novel (N) classes. For simplicity, we use mAP0.5 for base classes and nAP0.5 for novel classes, respectively.

Experimental settings. Further evaluations in this section show the mAP0.5 scores of baseline study BHRL [13], the proposed OSCDA framework (both with and without the CDQSS module enabled), and our other proposed method SACDA, on the 50 video sequences from the cross-video domain dataset VOT-LT 2019 [16].

For the evaluation of BHRL, the same experimental settings are used as in Section 4.2.

For the OSCDA method without CDQSS, the initial appearance of the target object is used as the query shot (Q), and the model is specifically fine-tuned with only this query shot (Q), once at the beginning of the inference phase. Following this, the model aims to identify all instances of the class of the query shot Q throughout the sequence.

In the case of evaluating the OSCDA method with CDQSS, the model is continuously fine-tuned during the inference on new query shots (Q) adaptively chosen by the CDQSS module as explained in Section 3.1. With each fine-tuning process, the model adapts to the latest appearance characteristics of the class of the query shot (Q), aiming to detect all instances of this class in the following frames.

At last, for the SACDA method, the model is fine-tuned once with the initial appearance of the target object and its variations (query shots Q) at the beginning of the inference phase. This approach aims for the model to adapt to different appearances of the target throughout the sequence without the need for continuous fine-tuning during the inference phase. Following this fine-tuning, the fine-tuned model aims to detect instances of the class of query shots (Q).

To evaluate these methods, the same evaluation strategy is employed. Upon completion of the inference process, the model's detections are then compared with the ground truth, and the mAP0.5 scores are calculated.

4.4.3 Performance analysis with COCO model

We first evaluate the proposed framework OSDCA by initiating the COCO model. As discussed before, we specifically initiate the baseline model pre-trained on the COCO split3. As discussed in Section 4.4.1, we treat each target object in each sequence as a different class and report the mAP0.5 scores for each sequence individually. We also group the video sequences as the base (*B*) and novel (*N*) according to target objects' main classes to report the performance across base and novel classes separately.

4.4.3.1 Performance analysis on base classes

Table 4.2 shows the mAP0.5 scores of the COCO model [13], the proposed OSCDA framework (both without and with the CDQSS module enabled), and our other proposed method SACDA, across the 50 video sequences from the cross-video domain dataset VOT-LT 2019 [16].

Table 4.2 : One-shot detection performance of the baseline and proposed methods on the VOT-LT 2019 dataset for the base classes of COCO split3 (mAP0.5).

Video Seq.	COCO model [13]	OSCDA (w/o CDQSS)	OSCDA (w/ CDQSS)	SACDA
ballet	**10.5**	42.9	36.6	**50.4**
bike1	**42.6**	66.3	**77.7**	72.8
dragon	**6.3**	31.0	39.3	**44.2**
group1	**17.7**	44.2	**51.5**	51.1
group2	**26.0**	45.3	35.8	**63.9**
group3	**57.0**	75.4	82.6	**88.0**
kitesurfing	58.2	76.5	**84.7**	**51.4**
longboard	75.0	81.1	**37.5**	**83.3**
tightrope	**22.1**	64.0	**71.4**	63.8
person2	87.5	92.9	**96.8**	**80.9**
person4	95.7	96.2	**97.5**	**95.1**
person5	100.0	100.0	100.0	100.0
person7	**94.5**	98.1	98.7	**99.0**
person14	**67.7**	79.0	**80.0**	88.7
person17	**92.5**	96.1	**98.7**	97.0
person19	**53.3**	70.4	**94.1**	66.9
person20	70.8	**68.6**	**98.6**	73.8
bicycle	**64.6**	**99.6**	84.5	77.7
wingsuit	**97.2**	98.7	**99.9**	98.8
skiing	**72.0**	88.0	**88.2**	84.6
rollerman	90.2	**88.3**	93.9	**94.8**
warmup	**24.6**	27.4	**38.3**	27.6
cat1	**47.0**	66.5	**82.4**	67.1
cat2	50.1	**62.0**	58.6	**49.5**
boat	75.8	**65.8**	**79.7**	69.2
freesbiedog	**71.6**	81.4	**83.2**	78.1
helicopter	**0.0**	**22.3**	21.6	**0.0**
bull	81.8	82.1	**90.2**	**74.7**
deer	**52.6**	66.9	**78.1**	67.8
dog	**28.1**	43.6	53.4	**59.7**
sitcom	**0.0**	68.3	**83.4**	67.1
uav1	24.9	**4.3**	**28.1**	9.2
horseride	14.2	**26.9**	26.6	**11.8**
yamaha	**86.8**	95.6	**97.2**	93.7
sup	**62.6**	69.0	**83.8**	74.2
freestyle	86.6	**93.0**	**70.8**	91.3
Avg.	**55.80**	68.51	**72.91**	68.81

As illustrated in Table 4.2, the proposed frameworks OSCDA (both without and with CDQSS) and SACDA significantly outperform the COCO model [13], with improvements of approximately 13%, 20%, and 14% in mAP0.5 scores across base classes, respectively. Among these, OSCDA with CDQSS enabled achieves the highest mAP0.5 scores overall but exhibits the lowest performance in two specific sequences: longboard and freestyle. In these sequences, both versions of OSCDA (without CDQSS) and SACDA demonstrate adequate and comparable performance, indicating that the models are effectively adapting to the appearance of the target object in the inference phase.

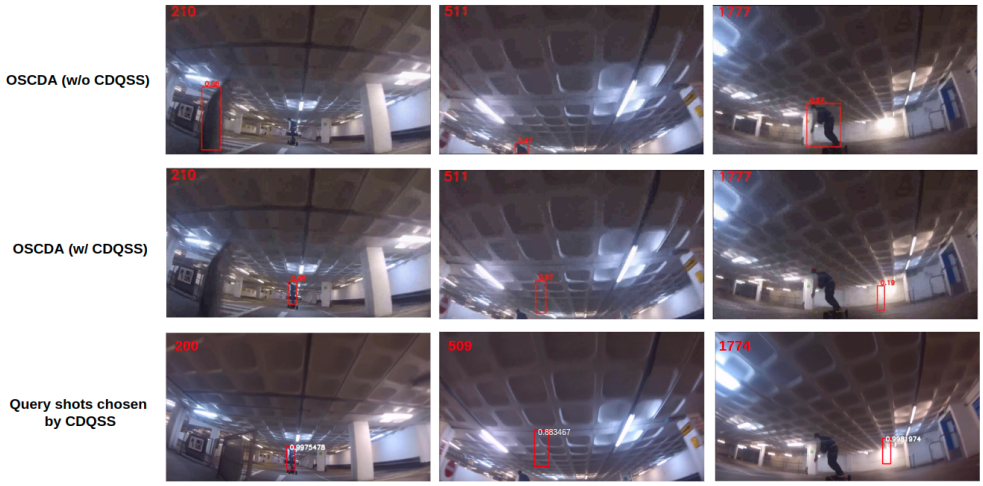

Figure 4.8 : Longboard-visual performance comparison between the OSCDA without and with CDQSS.

We investigate the performance decrease caused by enabling the CDQSS module by comparing the visual results and chosen query shots by CDQSS.

Figure 4.8 gives the visual comparison between the OSCDA framework without and with CDQSS with the updated query shots chosen by CDQSS.

As seen in 4.8, the detection performance of the model highly depends on the query shot selection capability of the CDQSS module. When the CDQSS accurately updates the query shot, as seen in the leftmost column, the model can successfully detect the target. However, if the CDQSS module fails to update the query shot effectively, the model fails to detect the region most similar to the incorrectly selected latest query shot, leading to errors as depicted in the second and third columns from the left.

Figure 4.9 : Group2-visual performance comparison between the COCO model, SACDA and OSCDA with CDQSS.

An interesting finding highlighted in Table 4.2 is that there are certain video sequences, ballet and group2, where the SACDA framework significantly outperforms the OSCDA (w/ CDQSS), deviating from the overall trend observed.

Figure 4.9 compares the visual results of the COCO model, SACDA, and OSCDA (w/ CDQSS).

As seen in Figure 4.9, this particular video sequence presents unique challenges inherent to the video domain, such as the presence of objects similar to the target and variations in the target's rotation and scale. However, the SACDA model is capable of adapting to these changes in scale and rotation across diverse video frames, even against shifting backgrounds. In contrast, the provided COCO model or the OSCDA (w/ CDQSS) framework can not show such a quick adaptation to the target in this specific scenario.

Table 4.3 : One-shot detection performance of the baseline and proposed methods on the VOT-LT 2019 dataset for the novel classes of COCO split3 (nAP0.5).

Video Seq.	COCO model [13]	OSCDA (w/o CDQSS)	OSCDA (w/ CDQSS)	SACDA
bird1	**0.0**	4.6	**10.1**	5.8
car1	**12.2**	42.4	**51.4**	39.2
car8	**69.0**	84.6	**99.9**	90.6
car9	**20.4**	30.2	**52.5**	47.2
car3	**9.5**	77.8	**83.9**	77.2
car6	**80.3**	85.4	**88.6**	83.9
car16	81.3	83.1	**99.6**	**71.7**
carchase	**3.2**	34.7	**31.1**	17.4
f1	**14.8**	86.0	91.1	**93.1**
following	82.3	89.3	**92.4**	**74.3**
liverRun	43.4	73.0	**81.8**	**7.1**
nissan	**50.7**	55.8	**73.7**	65.6
parachute	**61.5**	69.9	**84.6**	71.6
volkswagen	**66.3**	18.7	45.6	**44.0**
Avg.	**42.51**	59.67	**70.43**	56.33

4.4.3.2 Performance analysis on novel classes

Table 4.3 shows the nAP0.5 scores of the COCO model [13], the proposed OSCDA framework (both without and with the CDQSS module enabled), and our other proposed method SACDA, across the 50 video sequences from the cross-video domain dataset VOT-LT 2019 [16]. As illustrated in Table 4.3, the proposed frameworks OSCDA (both without and with CDQSS) and SACDA significantly outperform the COCO model [13], with improvements of approximately 17%, 28%, and 14% in mAP0.5 scores across novel classes, respectively. On the other hand, both the baseline model and the proposed methods show poorer nAP0.5 performance on the novel classes (N), compared to the one on the base classes in Table 4.3. This degradation is anticipated since the novel classes (N) are not included in the training phase, which typically leads to a decrease in performance for these classes.

The proposed OSCDA with CDQSS significantly improves the BHLR's performance on two specific sequences, car3 and f1.

Figure 4.10 illustrates the visual performance comparison between the COCO model and the proposed OSCDA (w/ CDQSS) framework. As discussed before, in the video domain, the target desired to be detected in every frame of a video sequence is introduced at the beginning of the sequence. The main motivation of this thesis is to

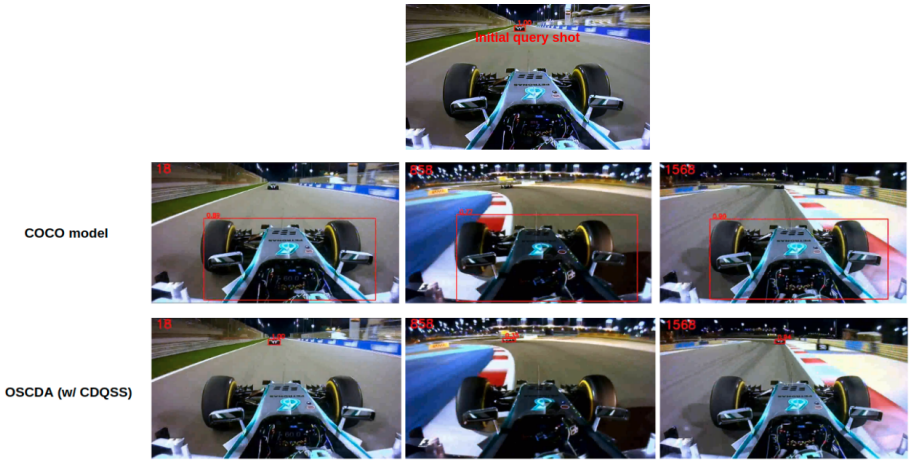

Figure 4.10 : F1-visual performance comparison between the COCO model and OSCDA with CDQSS.

enhance the baseline one-shot object detection model BHRL to enable BHRL to adapt to the appearance of target objects using just a single shot. Figure 4.10 demonstrates that the proposed OSCDA framework effectively adapts to the desired appearance of the target object and consistently detects it across video frames, despite rapid changes in the background, scene illumination, and the appearance of the target itself.

4.4.4 Performance analysis with VOC model

We evaluate the proposed framework OSDCA by initiating the VOC model. As discussed in Section 4.4.1, we treat each target object in each sequence as a different class and report the mAP0.5 scores for each sequence individually. We also group the video sequences as the base (B) and novel (N) according to target objects' main classes to report the performance across base and novel classes separately.

Table 4.4 : One-shot detection performance of the baseline and proposed methods on the VOT-LT 2019 dataset for the base classes of Pascal VOC (mAP0.5).

Video Seq.	VOC model [13]	OSCDA (w/o CDQSS)	OSCDA (w/ CDQSS)	SACDA
bird1	**0.0**	5.2	5.9	**11.7**
car1	**3.7**	20.3	19.3	**19.5**
car8	**71.1**	77.2	**94.8**	80.9
car9	**4.4**	48.9	**57.8**	44.3
car3	**0.2**	66.6	69.8	**75.1**
car6	**86.8**	88.1	**94.3**	88.2
car16	74.1	87.4	**95.4**	**64.4**
carchase	**1.2**	36.9	**49.0**	18.5
f1	**19.4**	94.1	**95.3**	94.0
following	87.0	86.8	**94.3**	88.9
liverRun	27.0	56.9	**67.8**	3.5
nissan	**55.6**	82.8	**84.6**	70.3
ballet	**12.3**	32.9	33.3	**56.8**
bike1	**55.2**	69.2	**83.4**	78.6
group1	**8.3**	33.5	42.6	**47.0**
group2	**9.7**	19.3	25.0	**29.5**
group3	**5.2**	43.0	54.7	**81.7**
kitesurfing	58.1	76.3	**79.0**	**52.9**
longboard	**62.3**	79.2	**80.9**	79.6
tightrope	**34.9**	**68.7**	64.4	67.3
person2	**66.6**	91.9	**97.4**	73.9
person4	**85.2**	97.7	**98.2**	97.7
person5	100.0	100.0	100.0	100.0
person7	97.0	92.5	**84.1**	**97.8**
person14	**15.6**	72.3	84.4	**94.2**
person17	**69.2**	97.0	**96.9**	**95.7**
person19	75.2	76.8	**80.1**	**74.2**
person20	64.6	**65.3**	**78.3**	68.7
bicycle	**74.8**	88.1	**89.7**	85.4
wingsuit	95.0	**91.6**	97.6	**98.4**
skiing	**66.5**	**89.2**	87.5	87.4
rollerman	90.4	**97.2**	97.7	96.5
warmup	**23.2**	40.7	**47.3**	41.2
boat	49.4	**59.6**	**65.3**	54.1
freesbiedog	**37.5**	78.3	**76.4**	71.4
dog	**10.0**	37.0	42.3	28.8
horseride	**4.9**	27.3	29.4	**52.3**
yamaha	**78.7**	93.2	**89.3**	89.1
sup	**29.8**	67.8	**83.8**	67.0
volkswagen	**48.3**	54.8	**64.0**	44.8
Avg.	**46.5**	67.29	**71.8**	66.8

4.4.4.1 Performance analysis on base classes

Table 4.4 shows the mAP0.5 scores of the VOC model [13], the proposed OSCDA framework (both without and with the CDQSS module enabled), and our other proposed method SACDA, across the 50 video sequences from the cross-video domain dataset VOT-LT 2019 [16].

As shown in Table 4.4, the proposed frameworks OSCDA (both without and with CDQSS) and SACDA significantly outperform the VOC model [13], with

Figure 4.11 : Bike1-visual performance comparison between the VOC model and OSCDA with CDQSS.

improvements of approximately 21%, 25%, and 20% in mAP0.5 scores across base classes, respectively.

Among all sequences, there is a significant performance improvement of almost 30% on the bike1 sequence between the VOC model and the proposed OSCDA (w/ CDQSS).

Figure 4.11 gives a visual performance comparison between the VOC model and the proposed OSCDA (w/ CDQSS) framework.

As shown in Figure 4.11, the proposed OSCDA framework demonstrates robustness against scale and orientation changes, as well as occlusions, which are significant challenges in the video domain. A major challenge in the video domain, particularly with the VOT-LT 2019 dataset, is the presence of multiple objects that closely resemble each other. This similarity often causes the model to make inaccurate detections. As seen in Figure 4.11, the proposed framework OSCDA overcomes these challenges, whereas the VOC model makes inaccurate detections.

Another interesting observation is that the SACDA method outperforms the OSCDA (w/ CDQSS) by 23% on ballet sequence. As discussed previously, the SACDA framework shows better performance on certain sequences in contrast to the overall trend.

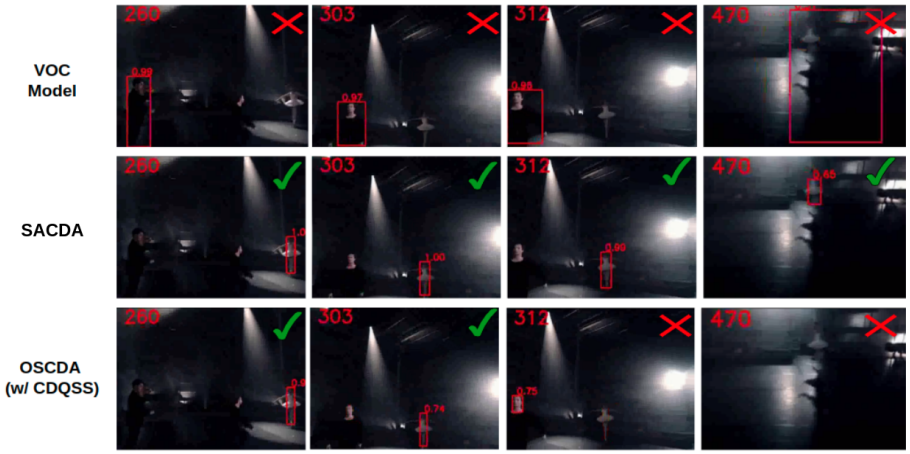

Figure 4.12 : Ballet-visual performance comparison between the VOC model and OSCDA with CDQSS.

Figure 4.12 compares the visual performance of the VOC model, SACDA and OSCDA (w/ CDQSS). As seen in Figure 4.12, the ballet sequence has notable challenges that are specific to the video domain, such as varying light exposures and significant changes in the target object's scale. Notably, in the second column of Figure 4.12, the OSCDA (w/ CDQSS) framework yields a lower confidence score compared to SACDA, even though it successfully detects the target object. For specific video sequences like ballet, group3, and horseride, the SACDA method consistently exhibits a more robust performance compared to OSCDA (w/ CDQSS).

Table 4.5 : One-shot detection performance of the baseline and proposed methods on the VOT-LT 2019 dataset for the novel classes of VOC (mAP0.5).

Video Seq.	VOC model [13]	OSCDA (w/o CDQSS)	OSCDA (w/ CDQSS)	SACDA
dragon	**9.1**	50.8	42.1	**56.6**
parachute	**18.5**	69.7	**75.2**	69.9
cat1	**46.3**	69.0	**87.1**	72.1
cat2	32.8	**59.08**	69.4	35.8
helicopter	**0.0**	**19.6**	19.5	**0.0**
bull	**69.3**	78.1	**82.4**	71.6
deer	**34.1**	54.2	**67.0**	55.2
sitcom	**0.0**	81.2	**83.4**	59.9
uav1	**10.2**	25.2	**32.2**	24.9
freestyle	**1.5**	**88.6**	83.4	67.5
Avg.	**22.2**	59.62	**64.23**	51.3

4.4.4.2 Performance analysis on novel classes

Table 4.5 shows the nAP0.5 scores of the VOC model [13], the proposed OSCDA framework (both without and with the CDQSS module enabled), and our other proposed method SACDA, across the 50 video sequences from the cross-video domain dataset VOT-LT 2019 [16]. As illustrated in Table 4.5, the proposed frameworks OSCDA (both without and with CDQSS) and SACDA significantly outperform the COCO model [13], with improvements of approximately 37%, 42%, and 29% in mAP0.5 scores across novel classes, respectively.

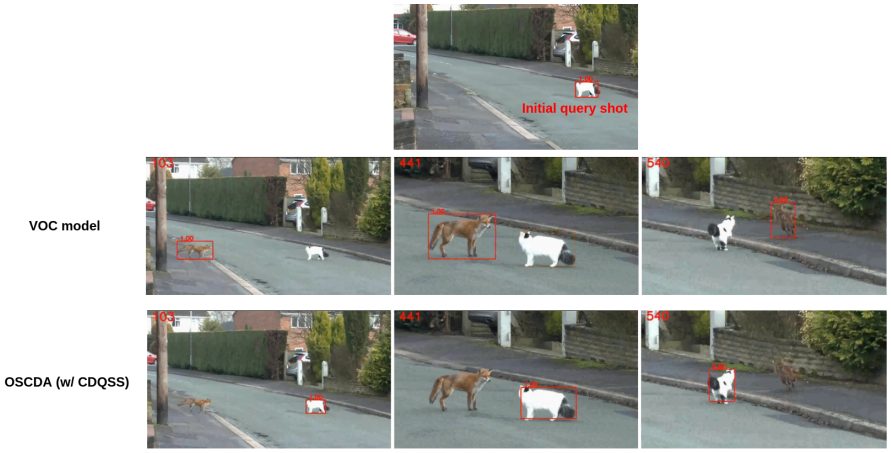

Figure 4.13 : Cat1-visual performance comparison between the VOC model and OSCDA with CDQSS.

One of the significant improvements is observed in the sequence cat1. Figure 4.13 gives a visual performance comparison between the VOC model and the proposed OSCDA (w/ CDQSS) framework. As seen in Figure 4.13, the video sequence presents common challenges associated with the video domain, such as the presence of similar objects in the scene and rapid changes in the scale and orientation of the target object. These challenges are typically absent in the still-image domain, resulting in the VOC model's inability to effectively handle them. In contrast, the proposed OSCDA framework enhances the VOC model's capability, making it robust to these types of challenges that arise due to the domain shift.

4.5 Extensive Analysis

This section provides extensive analysis of the experiments and results. We first evaluate the model performance on the VOT-LT 2019 dataset [16] without excluding the frames in which the target object is absent. While this is an unconventional approach for evaluating object detection models, it provides valuable insights into the model's performance in such scenarios, which is useful for future studies. For clarification, we refer to these scenarios as "full sequences", whereas calling "filtered sequences" to the ones we exclude the frames in which the target does not exist. in the following sections. Following this analysis, we report the average number of query shot updates performed with the proposed framework OSCDA (w/ CDQSS). Subsequently, we delve into the online fine-tuning process, which is the cornerstone of our proposed methods. Finally, we provide a visual comparison between the baseline model [13] and the proposed framework OSCDA (w/ CDQSS).

4.5.1 Evaluation on full video sequences

Before evaluations have been conducted on the VOT-LT 2019 video sequences by excluding the frames in which the target object is absent. We intentionally chose this reporting method because OSOD model evaluations are typically conducted on images where the target object is present. However, real video sequences include frames without the target object, and we compare the performance of proposed methods on sequences that include these frames as well. In this section, we divide the sequences into two categories: those with target object exits-enters and those without. We evaluate the proposed methods in both categories. Table 4.6 shows the mAP0.5 results of the methods on the sequences with target object exits-enters. As seen in Table 4.6, the OSCDA (w/ CDQSS) method shows a significant decrease compared to the OSCDA (w/o CDQSS) on the sequences with target object exits-enters.

Table 4.6 : Performance comparison on the sequences with target object exits-enters (mAP0.5).

Video Seq.	COCO Model	OSCDA (w/o CDQSS)	OSCDA (w/ CDQSS)	SACDA
ballet	**10.5**	42.9	27.8	**50.4**
dragon	**6.3**	31.0	21.9	**44.2**
group2	**26.0**	45.3	45.5	**63.9**
group3	**57.0**	75.4	66.8	**88.8**
longboard	75.0	81.1	**39.2**	**83.3**
tightrope	**22.1**	**64.0**	26.9	63.8
person7	94.5	98.1	**82.4**	**99.0**
person14	**67.7**	79.0	68.4	**88.7**
person17	92.5	96.1	**70.6**	**97.0**
person19	**53.3**	**70.4**	68.7	66.9
bicycle	64.6	**99.6**	**64.1**	77.7
wingsuit	97.2	98.7	**69.3**	**98.8**
skiing	72.0	**88.0**	**49.2**	84.6
rollerman	90.2	88.3	**43.5**	**94.8**
warmup	**24.6**	27.4	**30.9**	27.6
cat1	47.0	66.5	**34.2**	**67.1**
cat2	50.1	**62.0**	**49.3**	49.5
boat	**75.8**	65.8	**52.0**	69.2
freesbiedog	71.6	**81.4**	**62.6**	78.1
helicopter	**0.0**	22.3	**35.8**	**0.0**
bull	81.8	**82.1**	**55.4**	74.7
deer	52.6	66.9	**29.7**	**67.8**
dog	**28.1**	43.6	31.1	**59.7**
sitcom	**0.0**	**68.3**	36.3	67.1
uav1	**24.9**	**4.3**	21.8	9.2
horseride	14.2	26.9	**28.8**	**11.8**
yamaha	86.8	**95.6**	**41.7**	93.7
sup	62.6	69.0	**61.1**	**74.2**
freestyle	86.6	**93.0**	**53.9**	91.3
bird1	**0.0**	4.6	2.7	**5.8**
car1	**12.2**	42.4	**42.7**	39.2
carchase	**3.2**	**34.7**	21.7	17.4
f1	**14.8**	86.0	37.9	**93.1**
following	82.3	**89.3**	**40.0**	74.3
liverRun	43.4	**73.0**	30.3	**7.1**
nissan	**50.7**	55.8	53.4	**65.6**
parachute	61.5	69.9	**43.0**	**71.6**
volkswagen	**66.3**	**18.7**	31.6	44.0
Avg.	49.21	**63.35**	**44.01**	62.13

Table 4.7 : Performance comparison on the sequences without target object exits-enters (mAP0.5).

Video Seq.	COCO Model	OSCDA (w/o CDQSS)	OSCDA (w/ CDQSS)	SACDA
bike1	**42.6**	66.3	**77.7**	72.8
group1	17.7	44.2	**51.5**	**5.1**
car16	81.3	83.1	**99.6**	**71.7**
car9	**20.4**	30.2	**52.5**	47.2
car8	**69.0**	84.6	**99.9**	90.6
car6	**80.3**	85.4	**88.6**	83.9
car3	**9.5**	77.8	**83.9**	77.2
person20	70.8	**68.6**	**98.6**	73.8
person5	100.0	100.0	100.0	100.0
person4	95.7	96.2	**97.5**	**95.1**
person2	87.5	92.9	**96.8**	**80.9**
kitesurfing	58.2	76.5	**84.7**	**51.4**
Avg.	**61.08**	75.48	**85.94**	70.81

Table 4.7 compares the proposed methods on the sequences in which there the target object never leaves the scene. As illustrated in Figures 4.6 and 4.7, the mAP0.5 performance of the OSCDA (w/ CDQSS) method is almost reduced by half when evaluated on sequences where the target object exits and re-enters the scene, compared to the ones where the target remains in the scene continuously. This is an expected result since including the frames in which the target object does not exist in the evaluation may lead the CDQSS module to choose one of these frames as the updated query. In such scenarios, the model adapts to this appearance through online unsupervised fine-tuning, leading to continued incorrect predictions in subsequent frames.

Figure 4.14 illustrates the queries updated by CDQSS on full and filtered sequences. As seen in Figure 4.14, the CDQSS module highly tends to update the query shot incorrectly when using the full sequences compared to the filtered sequences.

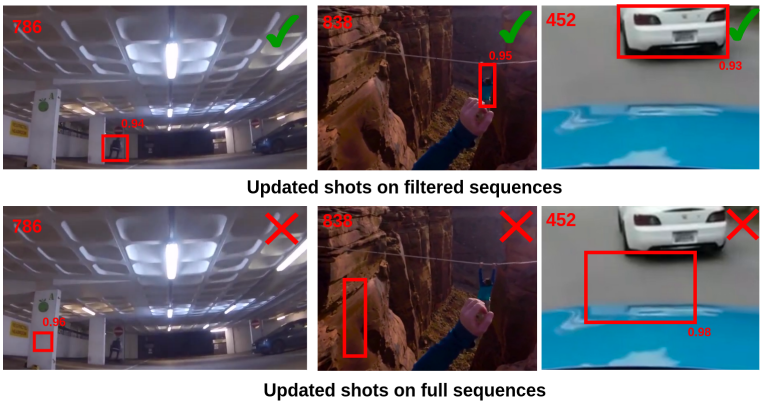

Figure 4.14 : Comparison between the CDQSS query updates on full and filtered sequences

4.5.2 Target updates

Table 4.8 illustrates the number of query shot updates performed by the CDQSS module with the total number of video frames in each video sequence from the VOT-LT 2019 dataset [16]. As expected, the frequency of query shot updates rises in conjunction with an increase in the total number of video frames in the sequence. The average number of query shots performed by the CDQSS is approximately 27, which is a quite small number considering the average number of video frames across all sequences is 3688. This result is encouraging as the number of query shot updates has a direct impact on the computational cost associated with the fine-tuning process. The OSCDA (w/ CDQSS) module carries out continuous unsupervised fine-tuning using these updated query shots throughout the video sequence, making efficiency crucial.

Table 4.8 : Number of shot updates across video sequences.

Seq.	# Query shot update	# Return to initial query shot	Seq.	# Query shot update	# Return to initial query shot
ballet (1127 fr.)	7	3	cat1 (1100 fr.)	9	1
bike1 (3085 fr.)	28	1	cat2 (2576 fr.)	19	5
dragon (3588 fr.)	19	15	bird1 (1380 fr.)	5	7
group1 (4873 fr.)	20	12	boat (5403 fr.)	33	20
group2 (2475 fr.)	35	10	freesbiedog (2990 fr.)	21	7
group3 (5322 fr.)	43	9	helicopter (708 fr.)	3	3
kitesurfing (4665 fr.)	37	7	car1 (2424 fr.)	13	10
longboard (5489 fr.)	76	23	car8 (2575 fr.)	23	1
tightrope (1889 fr.)	14	3	car9 (1879 fr.)	10	7
person2 (2623 fr.)	24	1	car3 (1717 fr.)	14	2
person4 (2743 fr.)	25	1	car6 (4861 fr.)	46	1
person5 (2101 fr.)	19	1	car16 (1993 fr.)	17	1
person7 (1995 fr.)	17	1	bull (1856 fr.)	16	1
person14 (2885 fr.)	25	23	carchase (8660 fr.)	78	7
person17 (2289 fr.)	20	1	deer (1343 fr.)	10	2
person19 (4006 fr.)	37	2	dog (1678 fr.)	8	7
bicycle (2280 fr.)	19	2	f1 (2786 fr.)	20	6
person20 (1783 fr.)	15	1	following (9031 fr.)	76	13
wingsuit (2172 fr.)	19	1	freestyle (1666 fr.)	14	1
skiing (2172 fr.)	18	2	liverRun (26277 fr.)	174	34
sup (2797 fr.)	22	4	nissan (3634 fr.)	30	5
parachute (2578 fr.)	21	3	sitcom (3119 fr.)	25	5
rollerman (1221 fr.)	10	1	uav1 (3252 fr.)	15	16
warmup (3681 fr.)	18	17	volkswagen (5141 fr.)	31	19
yamaha (2304 fr.)	19	3	horseride (14249 fr.)	77	64

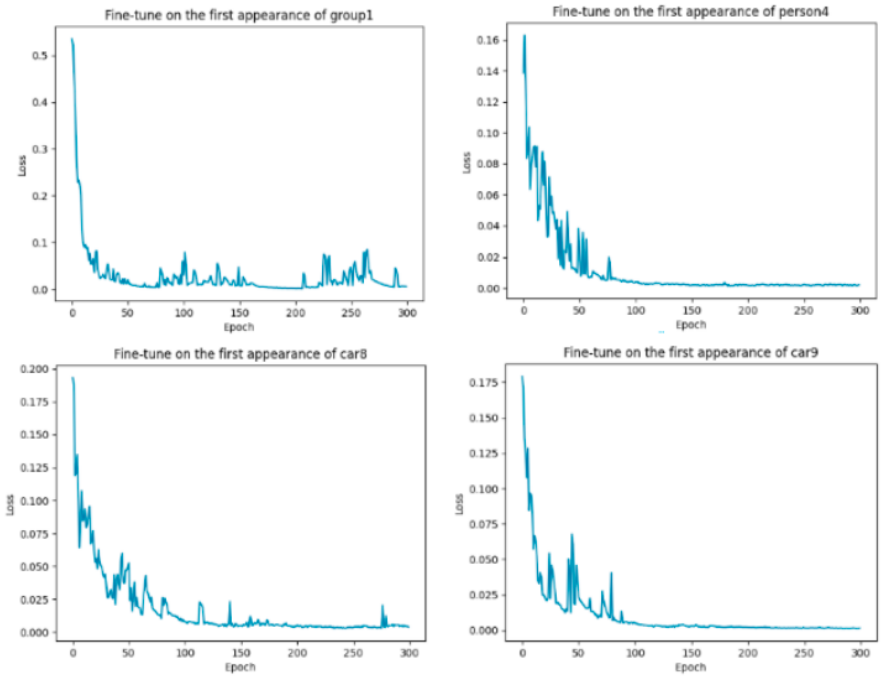

Figure 4.15 : Losses.

4.5.3 Detail analysis of fine-tuning

In this section, we provide a detailed analysis of the proposed online unsupervised fine-tuning process performed by the OSCDA (w/o CDQSS) module. For all the fine-tuning experiments, we fine-tune the COCO model [13] with a batch size of 1 on a 1 GPU using the SGD optimizer with a learning rate that starts at 0.02 for 100 epochs.

Figure 4.15 illustrates the change in loss functions during the online fine-tuning, initiated by the COCO model, which is performed with the initial query shot across 300 epochs.

As seen in Figure 4.15, the loss change is balanced after 100 epochs for both base B (group1 and person4) and novel N (car8 and car9) classes. These results encouraged us to fine-tune the provided models during 100 epochs in the experiments explained in previous sections.

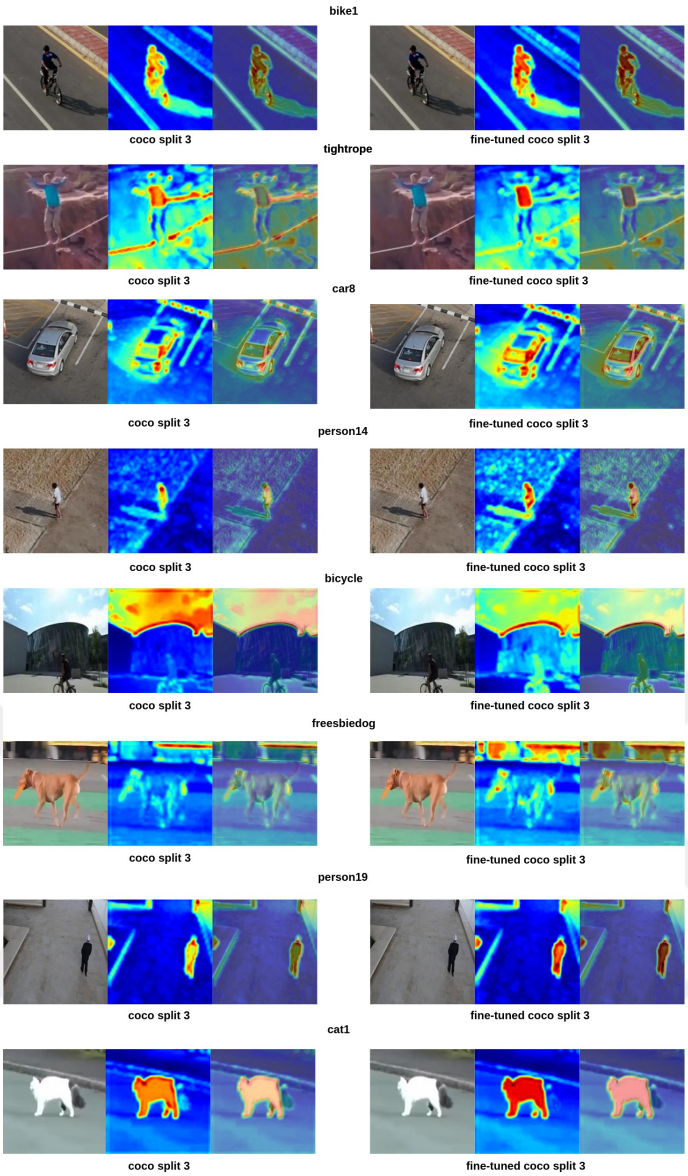

Figure 4.16 : Heatmaps.

Heatmaps. We employ heatmaps to analyze and visualize the effectiveness of the fine-tuning process, performed by initiating the COCO model, conducted with the initial query shot of the corresponding sequence. By comparing heatmaps generated before and after the fine-tuning process, we evaluate how the fine-tuning impacted the model's performance across different variables.

Figure 4.16 shows the heatmaps generated by the COCO model (coco split3 on the left) and those created by the fine-tuned COCO model (fine-tuned coco split3 on the right). These visual representations were based on the output scores of the model. Red regions indicate the parts where the model has high confidence in object detection,

whereas the blue regions denote areas of low confidence. In this scope, in Figure 4.16, regions that turned from blue to red represent the regions where the model's sensitivity to detecting objects has improved.

As seen in Figure 4.16, in both base and novel classes, the model begins to focus more on the target object following the fine-tuning process, indicating that it detects the target object with a higher confidence score.

4.5.4 Visual results

In this section, we compare the visual performance of the COCO model and the proposed framework OSCDA (w/ CDQSS) on challenging video sequences from the VOT-LT 2019 dataset [16]. Figures 4.17, 4.18, 4.19 show the overall performance comparison between the COCO model and the proposed OSCDA framework (w/ CDQSS).

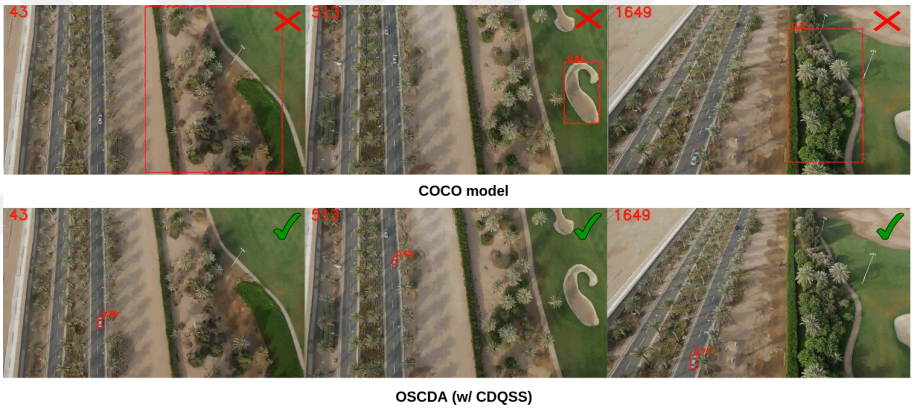

Figure 4.17 : Car3-visual comparison between the COCO model and the proposed OSCDA (w/ CDQSS).

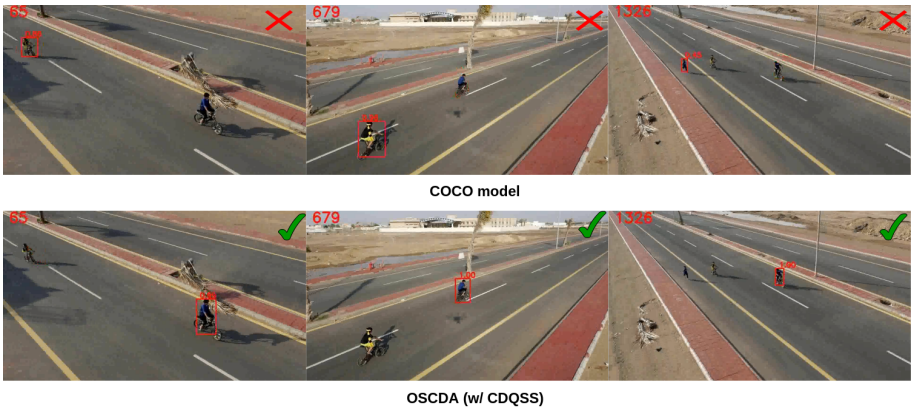

Figure 4.18 : Bike1-visual comparison between the COCO model and the proposed OSCDA (w/ CDQSS).

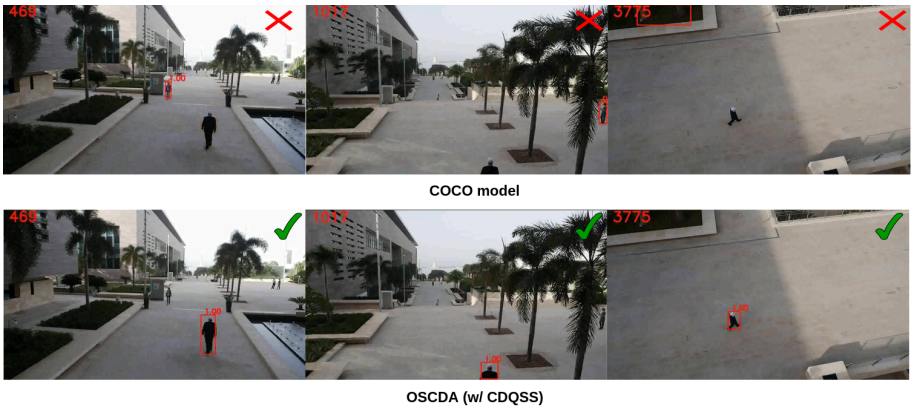

Figure 4.19 : Person19-visual comparison between the COCO model and the proposed OSCDA (w/ CDQSS).

5. CONCLUSIONS

In this thesis, our aim was to alleviate the performance degradation experienced by OSOD models during cross-domain evaluation. Unlike previous studies, our research specifically analyzes the domain shift challenges occurring between still images and video domains, rather than across diverse still image domains. Initially, we demonstrate the performance gap that OSOD models encounter in cross-domain settings. Subsequently, we introduce a novel one-shot object detector, OSCDA, which improves upon the existing OSOD model, BHRL [13], by incorporating an online fine-tuning technique. The proposed method OSCDA alleviates the performance degradation suffered by most one-shot object detectors due to domain shifts between still images and video data. Our proposed method OSCDA achieves SOTA performance across both novel and base classes within the video dataset VOT-LT [16], effectively overcoming the challenges of domain shift in one-shot object detection tasks. The proposed OSCDA method enhances the baseline BHRL model pre-trained on the COCO dataset [15] by 17% and 28% in mAP0.5 scores on base and novel classes, respectively. Moreover, the OSCDA outperforms the BHRL model pre-trained on the Pascal VOC dataset [17] by 25% and 42% in mAP0.5 scores on base and novel classes, respectively. These improvements demonstrate OSCDA's robustness and effectiveness in addressing domain shifts in one-shot object detection across still image and video domains.

On the other hand, the proposed method OSCDA (w/ CDQSS) could not handle the scenarios in which the target object is absent. Although such scenarios conventionally are not included in the evaluation phase of the OSOD models, addressing this issue is essential for enhancing the robustness of OSOD. Moreover, if this weakness can be

handled, this study could serve as a foundation for transforming the baseline model into a tracker in future research.

REFERENCES

[1] **Michaelis, C.**, **Ustyuzhaninov, I.**, **Bethge, M. and Ecker, A.S.** (2018). One-Shot Instance Segmentation, *CoRR*, *abs/1811.11507*, http://arxiv.org/abs/1811.11507, 1811.11507.

[2] **Guo, Y.**, **Li, Y.**, **Feris, R.S.**, **Wang, L. and Rosing, T.** (2019). Depthwise Convolution is All You Need for Learning Multiple Visual Domains, *CoRR*, *abs/1902.00927*, http://arxiv.org/abs/1902.00927, 1902.00927.

[3] **Hassan, E.**, **Khalil, Y. and Ahmad, I.** (2020). Learning Feature Fusion in Deep Learning-Based Object Detector, *Journal of Engineering*, *2020*, 1–11.

[4] **San Biagio, M.**, **Martelli, S.**, **Crocco, M.**, **Cristani, M. and Murino, V.** (2014). Encoding structural similarity by cross-covariance tensors for image classification, *International Journal of Pattern Recognition and Artificial Intelligence*, *28*, 19.

[5] **Wang, X.**, **Huang, T.E.**, **Darrell, T.**, **Gonzalez, J.E. and Yu, F.** (2020). *Frustratingly Simple Few-Shot Object Detection*, 2003.06957.

[6] **Wu, J.**, **Liu, S.**, **Huang, D. and Wang, Y.** (2020). *Multi-Scale Positive Sample Refinement for Few-Shot Object Detection*, 2007.09384.

[7] **Sun, B.**, **Li, B.**, **Cai, S.**, **Yuan, Y. and Zhang, C.** (2021). *FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding*, 2103.05950.

[8] **Qiao, L.**, **Zhao, Y.**, **Li, Z.**, **Qiu, X.**, **Wu, J. and Zhang, C.** (2021). DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection, *CoRR*, *abs/2108.09017*, https://arxiv.org/abs/2108.09017, 2108.09017.

[9] **Fu, Y.**, **Wang, Y.**, **Pan, Y.**, **Huai, L.**, **Qiu, X.**, **Shangguan, Z.**, **Liu, T.**, **Fu, Y.**, **Gool, L.V. and Jiang, X.** (2024). *Cross-Domain Few-Shot Object Detection via Enhanced Open-Set Object Detector*, 2402.03094.

[10] **Hsieh, T.**, **Lo, Y.**, **Chen, H. and Liu, T.** (2019). One-Shot Object Detection with Co-Attention and Co-Excitation, *CoRR*, *abs/1911.12529*, http://arxiv.org/abs/1911.12529, 1911.12529.

[11] **Chen, D.J.**, **Hsieh, H.Y. and Liu, T.L.** (2021). Adaptive Image Transformer for One-Shot Object Detection, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.12242–12251.

[12] **Fan, Q.**, **Zhuo, W. and Tai, Y.** (2019). Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector, *CoRR*, *abs/1908.01998*, http://arxiv.org/abs/1908.01998, 1908.01998.

[13] **Yang, H.**, **Cai, S.**, **Sheng, H.**, **Deng, B.**, **Huang, J.**, **Hua, X.**, **Tang, Y. and Zhang, Y.** (2022). Balanced and Hierarchical Relation Learning for One-shot Object Detection, *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, IEEE, pp.7581–7590, https://doi.org/10.1109/CVPR52688.2022.00744.

[14] **Kalogeiton, V.**, **Ferrari, V. and Schmid, C.** (2015). Analysing domain shift factors between videos and images for object detection, *CoRR*, *abs/1501.01186*, http://arxiv.org/abs/1501.01186, 1501.01186.

[15] **Lin, T.**, **Maire, M.**, **Belongie, S.J.**, **Bourdev, L.D.**, **Girshick, R.B.**, **Hays, J.**, **Perona, P.**, **Ramanan, D.**, **Dollár, P. and Zitnick, C.L.** (2014). Microsoft COCO: Common Objects in Context, *CoRR*, *abs/1405.0312*, http://arxiv.org/abs/1405.0312, 1405.0312.

[16] **Kristan, M.e.a.** (2019). The Seventh Visual Object Tracking VOT2019 Challenge Results, *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp.2206–2241.

[17] **Everingham, M.**, **Gool, L.V.**, **Williams, C.K.I.**, **Winn, J.M. and Zisserman, A.** (2010). The Pascal Visual Object Classes (VOC) Challenge., *Int. J. Comput. Vis.*, *88*(2), 303–338, http://dblp.uni-trier.de/db/journals/ijcv/ijcv88.html#EveringhamGWWZ10.

[18] **He, K.**, **Gkioxari, G.**, **Dollár, P. and Girshick, R.B.** (2017). Mask R-CNN, *CoRR*, *abs/1703.06870*, http://arxiv.org/abs/1703.06870, 1703.06870.

[19] **Ren, S.**, **He, K.**, **Girshick, R.B. and Sun, J.** (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *CoRR*, *abs/1506.01497*, http://arxiv.org/abs/1506.01497, 1506.01497.

[20] **Guo, Y.**, **Codella, N.C.F.**, **Karlinsky, L.**, **Smith, J.R.**, **Rosing, T. and Feris, R.S.** (2019). A New Benchmark for Evaluation of Cross-Domain Few-Shot Learning, *CoRR*, *abs/1912.07200*, http://arxiv.org/abs/1912.07200, 1912.07200.

[21] **Lee, K.**, **Yang, H.**, **Chakraborty, S.**, **Cai, Z.**, **Swaminathan, G.**, **Ravichandran, A. and Dabeer, O.** (2022). *Rethinking Few-Shot Object Detection on a Multi-Domain Benchmark*, 2207.11169.

[22] **Cai, J.**, **Li, C.**, **Tao, X.**, **Yuan, C. and Tai, Y.W.** (2022). DeViT: Deformed Vision Transformers in Video Inpainting, *Proceedings of the 30th ACM*

International Conference on Multimedia, MM '22, ACM, http://dx.doi.org/10.1145/3503161.3548395.

[23] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2015). Deep Residual Learning for Image Recognition, *CoRR*, *abs/1512.03385*, http://arxiv.org/abs/1512.03385, 1512.03385.

[24] **Lin, T.**, **Dollár, P.**, **Girshick, R.B.**, **He, K.**, **Hariharan, B. and Belongie, S.J.** (2016). Feature Pyramid Networks for Object Detection, *CoRR*, *abs/1612.03144*, http://arxiv.org/abs/1612.03144, 1612.03144.

[25] **Li, B.**, **Wu, W.**, **Wang, Q.**, **Zhang, F.**, **Xing, J. and Yan, J.** (2018). SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks, *CoRR*, *abs/1812.11703*, http://arxiv.org/abs/1812.11703, 1812.11703.

[26] **Chollet, F.** (2016). Xception: Deep Learning with Depthwise Separable Convolutions, *CoRR*, *abs/1610.02357*, http://arxiv.org/abs/1610.02357, 1610.02357.

[27] **Radford, A.**, **Kim, J.W.**, **Hallacy, C.**, **Ramesh, A.**, **Goh, G.**, **Agarwal, S.**, **Sastry, G.**, **Askell, A.**, **Mishkin, P.**, **Clark, J.**, **Krueger, G. and Sutskever, I.** (2021). Learning Transferable Visual Models From Natural Language Supervision, *CoRR*, *abs/2103.00020*, https://arxiv.org/abs/2103.00020, 2103.00020.

[28] **OpenAI** (2023). *ChatGPT-4*, https://openai.com/chatgpt, accessed: INSERT-ACCESS-DATE-HERE.

CURRICULUM VITAE

Name SURNAME: İrem Beyza ONUR

EDUCATION:

- **B.Sc.:** 2021, Istanbul Technical University, Faculty of Electrical and Electronics Engineering, Electronics and Communication Engineering Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2021-Cont. Autonomous Vehicle Engineer at Ford Otosan.

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **İ. B. Onur.**, F. Gurkan and B. Gunsel, "Video Object Verification via Meta-learning," 2022 30th Signal Processing and Communications Applications Conference (SIU), Safranbolu, Turkey, 2022, pp. 1-4, doi: 10.1109/SIU55565.2022.9864850.