

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**DOMAIN GENERALIZED OBJECT DETECTION AND
INSTANCE SEGMENTATION FOR REMOTE SENSING IMAGES**

EFKAN DURAKLI

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASSIST. PROF. DR. YAKUP GENÇ

JULY 2024

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**DOMAIN GENERALIZED OBJECT DETECTION AND
INSTANCE SEGMENTATION FOR REMOTE SENSING
IMAGES**

EFKAN DURAKLI

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASSIST. PROF. DR. YAKUP GENÇ

Co-ADVISOR: ASSOC. PROF. DR. ERCHAN APTOULA

JULY 2024

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

UZAKTAN ALGILAMA GÖRÜNTÜLERİ İÇİN ALAN
GENELLEŞTİRİLMİŞ NESNE TANIMA VE ÖRNEK
BÖLÜTLEME

EFKAN DURAKLI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: DR. ÖĞR. ÜYESİ YAKUP GENÇ

II. DANIŞMAN: DOÇ. DR. ERCHAN APTOULA

TEMMUZ 2024



MASTER of SCIENCE JURY APPROVAL FORM

A thesis submitted by Efkan Duraklı defended on 16/07/2024 before the jury formed with the 01/07/2024 date and 2024/33 numbered decision of the GTU Graduate Administration Board, has been accepted as a MASTER of SCIENCE thesis in the Department of Computer Engineering.

JURY

MEMBER (THESIS ADVISOR) : Assist. Prof. Dr. Yakup GENÇ

MEMBER (II. ADVISOR) : Assoc. Prof. Dr. Erchan APTOULA

MEMBER : Assoc. Prof. Dr. Mehmet GÖKTÜRK

MEMBER : Assoc. Prof. Dr. Habil KALKAN

MEMBER : Assoc. Prof. Dr. Alp ERTÜRK

APPROVAL

Gebze Technical University Graduate Administration Board
...../...../..... date and/..... numbered decision.

SIGNATURE/SEAL

ABSTRACT

Building roof type detection and building instance segmentation are crucial tasks for remote sensing applications such as urban planning and disaster management. Recent advances in deep learning-based object detection and segmentation have demonstrated outstanding performance in these fields. However, domain shift is a significant challenge, where differences between training and test data distributions can lead to performance degradation. Traditional domain adaptation methods require access to target domain data, which is only sometimes feasible. This thesis proposes novel domain generalization methods that address domain shift without requiring access to target data. The proposed approach tackles domain shift at both the instance and image levels for roof-type detection from remote sensing images. Consistency regularization is imposed to enforce alignment between image-level and instance-level domain predictions. The proposed method, representing a pioneering effort in domain generalization for remote sensing object detection, is evaluated using the IEEE Data Fusion Contest 2023 dataset. Similarly, for building instance segmentation, Mask R-CNN is enhanced with image, instance, and pixel-level domain adversarial modules to extract domain-invariant features, aiming to solve the problem in a domain generalization setting without target labels or data. Promising outcomes are shown by preliminary results from cross-continent domain generalization experiments, indicating the potential of proposed methods to improve the robustness of remote sensing applications in varied contexts.

Keywords: Remote Sensing, Object Detection, Instance Segmentation, Domain Generalization, Building Detection, Building Segmentation

ÖZET

Çatı tipi tespiti ve bina örnekleme ayrımı, kentsel planlama ve afet yönetimi gibi uzaktan algılama uygulamaları için önemli görevlerdir. Derin öğrenme tabanlı nesne tespiti ve ayrımı konusundaki son gelişmeler, bu alanlarda olağanüstü performans sergilemiştir. Ancak, önemli bir zorluk, eğitim ve sınama veri dağılımları arasındaki farkların performans düşüşüne yol açabileceği alan kaymasıdır. Geleneksel alan uyarlama yöntemleri hedef alan verilerine erişim gerektirir ki bu her zaman mümkün değildir.

Bu bağlamda, bu tez, hedef alan verilerine erişim gerektirmeyen alan kaymasını ele alan yeni alan genelleme yöntemleri önermektedir. Uzaktan algılama görüntülerinden çatı tipi tespiti için önerilen yaklaşım, hem örnek hem de görüntü seviyesinde alan kaymasını ele alır. Ayrıca, görüntü ve örnek seviyesindeki alan tahminleri arasında hizalamayı sağlamak için tutarlılık düzenlemesi uygulanır. Uzaktan algılama nesne tespiti için alan genelleme konusunda öncü bir çaba olan önerilen yöntem, IEEE Data Fusion Contest 2023 veri seti kullanılarak değerlendirilmiştir.

Benzer şekilde, bina örnekleme ayrımı için, Mask R-CNN, hedef etiketleri veya verileri olmadan alan genelleme ayarında sorunu çözmeyi amaçlayarak alan bağımsız özellikler çıkarmak için görüntü, örnek ve piksel seviyesinde alan adversaryal modüllerle geliştirilmiştir. Kıtalar arası alan genelleme deneylerinden elde edilen ön sonuçlar, önerilen yöntemlerimizin farklı bağlamlarda uzaktan algılama uygulamalarının dayanıklılığını artırma potansiyelini göstererek ümit verici sonuçlar ortaya koymaktadır.

Anahtar Kelimeler: Uzaktan Algılama, Nesne Tespiti, Örnek Bölütleme, Alan Genelleme, Bina Tespiti, Bina Bölütleme

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors, Assoc. Prof. Erchan Aptoula and Assist. Prof. Yakup Genç, for their continuous support, guidance, and encouragement throughout my research and the writing of this thesis. Their invaluable insights and expertise have been instrumental in shaping this work.

I am very grateful for the support I received from my family: my brother Özkan, my father Zafer, and my mother Hatice, to whom I owe special thanks for her unwavering belief in me all the time.

A special mention goes to my wife, Yüksel Duraklı, for her unwavering support, patience, and encouragement. Her belief in me has been a constant source of motivation. Thank you all for your support and encouragement.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	v
ÖZET	vi
ACKNOWLEDGEMENT	vii
TABLE of CONTENTS	viii
LIST of ABBREVIATIONS and ACRONYMS	x
LIST of FIGURES	xi
LIST of TABLES	xii
1. INTRODUCTION	1
1.1. Problem Definition	1
1.2. Aim and Scope	2
1.3. Contributions	2
1.4. Outline	3
2. THEORETICAL BACKGROUND AND RELATED WORKS	4
2.1. Fundamentals of Remote Sensing	4
2.2. Object Detection and Instance Segmentation	6
2.2.1. Object Detectors	7
2.2.2. Instance Segmentation Approaches	12
2.2.3. Object Detection in Remote Sensing	15
2.3. Domain Generalization	16
2.3.1. Problem Definition	16
2.3.2. Related Research Areas	17
2.3.3. Domain Generalization Methods	18
2.4. Domain Generalization for Remote Sensing	21
3. DOMAIN GENERALIZED OBJECT DETECTION FOR REMOTE SENSING	23
3.1. Introduction	23
3.2. Preliminaries	25
3.3. Proposed Method	26
3.4. Experiments	28
3.4.1. Dataset	28
3.4.2. Experiment Setup	29
3.4.3. Results	31
3.4.4. Ablation Study	34
3.4.5. Conclusion	35

4. DOMAIN GENERALIZED MASK R-CNN FOR BUILDING INSTANCE SEGMENTATION	36
4.1. Introduction	36
4.2. Domain generalized Mask R-CNN	38
4.3. Experiments	40
4.3.1. Dataset	40
4.3.2. Experiment Setup	41
4.3.3. Results	42
4.3.4. Ablation Study	44
4.4. Conclusion	45
5. CONCLUSION	47
REFERENCES	49
BIOGRAPHY	60
PUBLICATIONS WITHIN the SCOPE of THESIS STUDY	61

LIST of ABBREVIATIONS and ACRONYMS

UAV	:	Unmanned Aerial Vehicle
CNN	:	Convolutional Neural Network
DL	:	Deep Learning
IEEE	:	Institute of Electrical and Electronics Engineers
DFC	:	Data Fusion Contest
HOG	:	Histogram of Gradients
DPM	:	Deformable Part-based Model
SVM	:	Support Vector Machine
R-CNN	:	Region-based Convolutional Neural Network
GRL	:	Gradient Reversal Layer
SGD	:	Stochastic Gradient Descent
IoU	:	Intersection over Union
FPN	:	Feature Pyramid Network
DNN	:	Deep Neural Network
NMS	:	Non-Maximum Suppression
RPN	:	Region Proposal Network
RoI	:	Region of Interest
ML	:	Machine Learning
DA	:	Domain Adaptation
DG	:	Domain Generalization
MTL	:	Multitask Learning
TL	:	Transfer Learning
SSD	:	Single Shot MultiBox Detector
DETR	:	DEtection TRansformers
HTC	:	Hybrid Task Cascade
mAP	:	Mean Average Precision
DGFRCNN-RS	:	Domain Generalized Faster R-CNN for Remote Sensing
DGMaskRCNN	:	Domain Generalized Mask R-CNN

LIST OF FIGURES

	<u>Page</u>
Figure 2.1: A visual overview of the remote sensing process [1].	4
Figure 2.2: An illustrative depiction of R-CNN [2].	8
Figure 2.3: A visual representation of Fast R-CNN [3].	9
Figure 2.4: A visual representation of Faster R-CNN [4].	10
Figure 2.5: A visual representation of YOLO [5].	11
Figure 2.6: Architecture of SSD [6].	11
Figure 2.7: A visual depiction of DETR [7].	12
Figure 2.8: A visual representation of Mask R-CNN [8].	14
Figure 2.9: Samples from the PACS [9] dataset for domain generalization. The training set comprises art paintings, cartoons, and sketch images, while the test set comprises photographic images. DG aims to learn from the training set to perform well on the unseen test set.	17
Figure 3.1: Buildings from different cities (DFC23 dataset); from top-left to bottom-right: Addis Ababa, New Delhi, Barcelona, Darwin (bottom-left), Brasilia, and New York.	24
Figure 3.2: An outline of the suggested Domain Generalized Faster R-CNN network, where RPN stands for region proposal network, Conv for convolutional layer, FC for fully connected layer, and GRL for gradient reversal layer.	27
Figure 4.1: Buildings from different cities (DFC23 dataset); from top-left to bottom-right: Addis Ababa, New Delhi, Barcelona, Darwin (bottom-left), Brasilia, and New York.	36
Figure 4.2: An outline of the proposed Domain Generalized Mask R-CNN model where FPN, RPN, Conv, FC, and GRL denote feature pyramid network, region proposal network, convolutional layer, fully connected layer, and gradient reversal layer, respectively.	37

LIST OF TABLES

	<u>Page</u>
Table 3.1: Distribution of images and instances across various cities on different continents	29
Table 3.2: Distribution of roof type categories in the IEEE DFC 2023 dataset, showing the number of instances and their respective percentages	30
Table 3.3: mAP and mAP at a 0.5 IoU threshold results on the IEEE DFC 2023 dataset for the two predominant categories: Flat and Gable. All other roof types are excluded from the dataset.	32
Table 3.4: Cross validation experiment results for building detection using different folds with the proposed DGFRCNN-RS model. Each fold used images from different continents for training, validation, and testing. The performance metrics reported include mAP and mAP at a 0.5 IoU threshold (mAP@50).	33
Table 3.5: Results of the ablation study for weights α , β , and γ in Equation 3.7, showing the impact on mean average precision (mAP) and mAP at a 0.5 IoU threshold (mAP@50) for the proposed DGFRCNN-RS model.	35
Table 4.1: Distribution of images and instances across various cities on different continents.	41
Table 4.2: Cross validation experiment results for building detection using different folds with the proposed DGMaskRCNN model. Each fold used images from different continents for training, validation, and testing. The performance metrics reported include mAP and mAP at a 0.5 IoU threshold (mAP@50).	43
Table 4.3 Results of the ablation study for weights α , β , and γ in Equation 4.5, showing the impact on mAP and mAP at a 0.5 IoU threshold (mAP@50) for the proposed DGMaskRCNN model.	45

1. INTRODUCTION

1.1. Problem Definition

The rapid progress in sensor technologies has greatly increased the quantity and quality of high-spatial-resolution remote sensing images. These advancements have enabled more sophisticated processing techniques like object detection and instance segmentation. Object detection identifies and classifies objects using bounding boxes, whereas instance segmentation goes further by producing pixel-level masks that accurately delineate the shape of each object instance. This ability to detect and segment objects with pixel-level precision is directly applicable to urban planning, such as identifying buildings and vehicles, and to maritime traffic monitoring for ship detection.

The advent of deep learning (DL) has resulted in convolutional neural networks (CNNs) showcasing exceptional performance in the domains of instance segmentation and object detection within remote sensing. Despite their effectiveness, these methods present notable challenges. They rely heavily on large amounts of labeled data, which is labor-intensive and time-consuming to gather. Furthermore, DL models presume that the training and testing datasets share the same distribution, an assumption that often fails in practical applications.

DL models usually struggle when there is a difference between the distributions of the test and training data, known as domain shift. Domain shift can occur due to various reasons, such as changes in lighting conditions, different viewpoints, variations in sensor types, and other environmental factors. Addressing domain shifts proves crucial for maintaining the resilience and precision of DL models when deployed in practical scenarios.

Two primary strategies for addressing domain shift are domain generalization (DG) and domain adaptation (DA).

DA entails modifying a model that has been trained on a source domain so that it can perform well on a distinct but related target domain. This adjustment process requires access to target domain data during the training phase.

Conversely, DG seeks to train a model using one or more source domains to ensure it

generalizes effectively to an unseen target domain without utilizing any target domain data during training. This is accomplished by learning features that remain consistent and reliable across various domains.

Both DA and DG address the issue of domain shift, but they employ fundamentally different approaches. DA leverages information from the target domain to adjust the model according to the specific attributes of the target setting. Conversely, DG strives to develop a universally resilient model capable of excelling in various domains.

1.2. Aim and Scope

This thesis aims to develop and assess new domain generalization methods specifically designed for remote sensing applications, with a particular emphasis on object detection and building instance segmentation. The specific objectives are:

Enhancement of Object Detection: A domain generalization method is developed to enhance the accuracy and reliability of object detection in remote sensing images. This approach tackles domain discrepancies at the image and object levels, incorporating a regularization method for consistency to reinforce robustness further.

Advancement of Building Instance Segmentation: To design a domain generalization approach to improve building instance segmentation. This method addresses domain shifts at image, instance, and pixel levels, ensuring more precise segmentation results.

Comprehensive Evaluation: To rigorously evaluate the proposed domain generalization methods using the IEEE (Institute of Electrical and Electronics Engineers) Data Fusion Contest 2023 dataset across various settings. This includes testing the methods under different conditions to validate their effectiveness and generalizability.

1.3. Contributions

The contributions of this thesis are the following:

- i) Developed a domain generalization method to enhance object detection in remote sensing images by addressing domain shifts at the image level and instance level. Furthermore, a method for enhancing the robustness of domain generalization

was proposed through the introduction of a consistency regularization technique.

- ii) Evaluated the proposed domain generalization method using the IEEE Data Fusion Contest (DFC) 2023 dataset [10] in two different settings.
- iii) Developed a domain generalization method to improve building instance segmentation in remote sensing images, addressing domain shifts at the image, instance, and pixel levels.
- iv) Assessed the proposed domain generalization method for building instance segmentation using the IEEE DFC 2023 dataset.
- v) The proposed domain generalization methods for object detection and instance segmentation are the first to apply domain generalization to these tasks in remote sensing images.

1.4. Outline

The thesis is organized as follows:

Chapter 2 provides the theoretical background necessary for understanding the research context, covering topics such as remote sensing, object detection, instance segmentation, and domain generalization.

Chapter 3 details the proposed domain generalization method specifically designed for object detection in remote sensing images.

Chapter 4 explains the domain generalization method for instance segmentation of buildings from remote sensing images, describing the enhancements made to the Mask R-CNN (Region-based Convolutional Neural Network) framework.

Chapter 5 concludes the thesis by providing a summary of the work and proposing directions for future research.

2. THEORETICAL BACKGROUND AND RELATED WORKS

2.1. Fundamentals of Remote Sensing

Remote sensing is a crucial process (Figure 2.1) in acquiring and interpreting information about the Earth's surface without physical contact. This process serves as the cornerstone of numerous scientific and practical applications, offering a broad spectrum of capabilities through the use of various platforms and sensors. The data collected via remote sensing can be harnessed for environmental monitoring [11], urban planning[12], disaster management [13], agriculture [14], and climate change studies [15], among other uses. Each platform used in remote sensing, including satellites [16], aircraft [17], and unmanned aerial vehicles(UAVs) [18], presents unique strengths and limitations, which are carefully considered depending on the specific requirements of a given application.

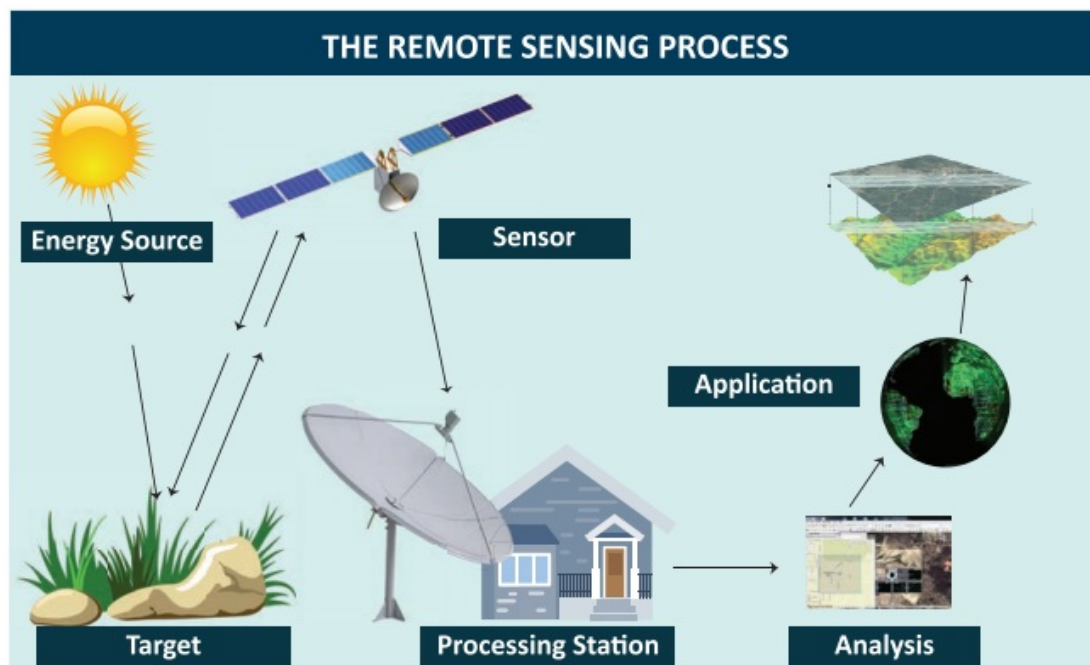


Figure 2.1: A visual overview of the remote sensing process [1].

Satellites, for example, provide extensive global coverage and the potential for long-

term temporal monitoring, making them particularly valuable for large-scale and longitudinal studies. However, the revisiting time of satellites can be prolonged, and their data may be subject to atmospheric interference, which can degrade image quality. Aircraft, on the other hand, offer higher spatial resolution and greater flexibility in data acquisition compared to satellites, making them ideal for regional studies and rapid response scenarios. The downside to using aircraft lies in their operational costs and the limited coverage they provide. UAVs, characterized by their high spatial and temporal resolution, are especially useful for detailed analysis of small areas. Nonetheless, their flight range, duration, and susceptibility to weather conditions impose certain operational constraints.

In addition to the diversity of platforms, the applications of remote sensing are equally varied. The technology is instrumental in environmental monitoring, where it is employed to track land-use changes, deforestation, and the health of water bodies, as well as to monitor natural disasters. In agriculture, remote sensing aids in assessing crop health, predicting yield potentials, and optimizing farming practices. Urban planning also benefits significantly from remote sensing, with applications ranging from the analysis of urban growth and infrastructure development to monitoring land cover changes. Disaster management is another critical area where remote sensing plays a crucial role, facilitating the mapping of flood zones, tracking wildfires, and evaluating damage following natural disasters. Moreover, remote sensing contributes to climate change studies by enabling the monitoring of climate patterns, sea-level changes, and the broader impacts of climate change.

However, the field of remote sensing is not without its challenges. One of the most significant hurdles is the sheer size and complexity of the data generated, which requires substantial computational resources for processing and analysis. Additionally, atmospheric conditions and other environmental factors can affect the quality of the captured data, introducing noise that must be accounted for during analysis. The limitations inherent to each sensor—whether in terms of spatial, spectral, or temporal resolution—necessitate careful selection based on the intended application to ensure the data's utility. Furthermore, the interpretation of remote sensing data is inherently complex, often demanding expertise in both image processing and the specific domain of application to extract meaningful insights.

The characteristics of remote sensing images further underscore the complexity of their analysis. These images are typically captured across multiple spectral bands, including those beyond the visible spectrum, such as infrared and microwave radiation, which can reveal otherwise invisible features and properties. The spatial resolution of remote sensing images varies, representing the Earth's surface with different degrees of detail that can range from meters to kilometers per pixel. Temporal resolution is another critical factor, providing data at different revisit frequencies, which is essential for monitoring changes over time. Additionally, the type of data generated by remote sensing systems can vary significantly, including optical imagery [19], radar imagery [20], hyperspectral imagery [21], and LiDAR data [22], each offering unique insights and applications.

In conclusion, the fundamentals of remote sensing encompass a wide array of platforms, applications, challenges, and image characteristics. This diversity highlights the importance of remote sensing as a tool for acquiring and interpreting critical information about the Earth's surface while also pointing to the ongoing need for advancements in technology and methodology to address the challenges faced by practitioners in the field.

2.2. Object Detection and Instance Segmentation

Object detection is a vital task in computer vision that entails identifying and classifying objects within an image. Unlike image classification, which assigns a singular label to an entire image, object detection predicts the object categories and their exact locations within the image. The primary aim of object detection is to provide a detailed understanding of an image's content, which is essential for many computer vision applications. This includes recognizing what objects are present and determining their positions and sizes.

Object detection is the foundation for more advanced tasks like scene understanding, instance segmentation, and object tracking. Instance segmentation, in particular, extends beyond traditional object detection by not only identifying and localizing objects within an image but also precisely segmenting each object. Both instance segmentation and object detection are crucial for a wide range of applications, from [23] remote sensing

applications such as urban planning [24, 25], ship detection [20, 26], and environmental monitoring [27, 28] to autonomous vehicles.

2.2.1. Object Detectors

The traditional object detection pipeline consists of three main stages: region selection, feature extraction, and classification [29]. During the region selection stage, methods like selective search [30] or sliding window approaches are used to identify potential areas within the image where objects might be found. These identified regions are then processed in the feature extraction stage, where handcrafted features such as scale-invariant feature transform (SIFT) [31], histogram of gradients (HOG) [32], and Haar-like features [33] are computed to capture the visual characteristics of the objects. In the final classification stage, these features are utilized to train ML classifiers such as support vector machine (SVM) [34], AdaBoost [35], deformable part-based model (DPM) [36], and decision trees to categorize the objects into predefined classes.

However, traditional object detection methods have several limitations. One major drawback is their reliance on handcrafted features and shallow architectures, which makes them less effective at handling scale variations, occlusions, and complex scenes.

Conversely, DL-based object detection methods offer significant advantages with the rise of deep neural networks (DNNs). They use CNNs to automatically learn discriminative features from data, enabling them to better manage complex object appearances and backgrounds. Two-stage and one-stage detectors are two types of DL-based object detection frameworks.

Two-stage Object Detectors

These types of object detectors follow the traditional object detection pipeline. As their name implies, object detection takes place in two stages. First, potential object regions are identified, and each proposal is classified into various object categories. Examples of detectors in this category include R-CNN [2], Fast R-CNN [3], Faster R-CNN [4], and Mask R-CNN [8].

R-CNN is the pioneering study that introduced the concept of using region proposals to detect objects within an image. It utilizes a technique called selective search to generate these region proposals. Subsequently, each proposal is warped into a fixed

size, as required by CNN, which is then employed to extract features, resulting in a final representation for each proposal. Thanks to their substantial learning capacity, expressive capabilities, and hierarchical structure, CNNs enable the extraction of detailed and reliable feature representations for each region. Following this, classifiers specific to each object category are applied across multiple classes, scoring different regions against positive and background regions. Subsequently, adjustments are made to these regions through bounding box regression, and they are filtered using a greedy nonmaximum suppression (NMS) technique, resulting in the final bounding boxes that preserve object locations. Fig. 2.2 shows an illustrative depiction of R-CNN.

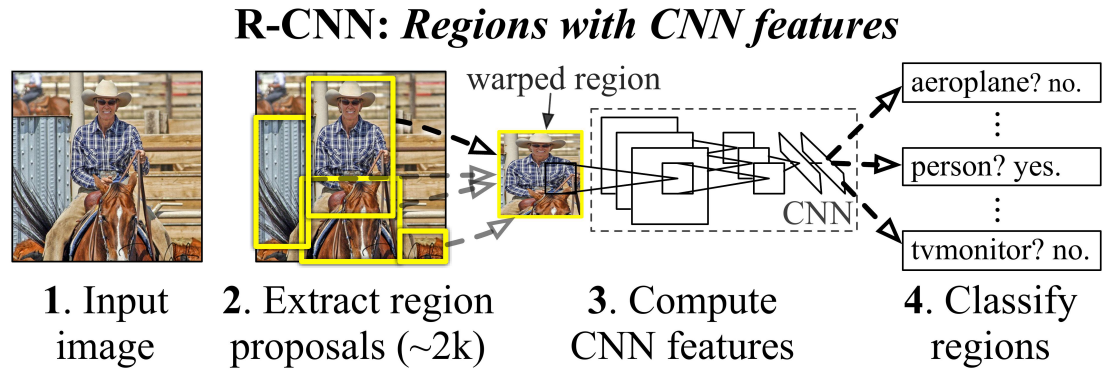


Figure 2.2: An illustrative depiction of R-CNN [2].

Fast R-CNN improves upon R-CNN by enhancing the efficiency of object detection. Unlike R-CNN, Fast R-CNN utilizes a shared CNN to efficiently extract features from the entire image instead of independently extracting CNN feature vectors for each region proposal. This streamlined approach removes the need to compute CNN features separately for each region proposal, significantly speeding up the process. It also replaces the pretrained categoric linear SVMs with a softmax layer. The introduction of the region of interest (RoI) pooling layer is crucial in aligning the region proposals with the CNN features, further enhancing the model's performance.

In summary, Fast R-CNN marks a substantial advancement in object detection, combining the accuracy of CNN-based feature extraction with the speed necessary for practical applications. Fig. 2.3 shows a visual representation of Fast R-CNN. Faster R-CNN builds upon the frameworks of R-CNN and Fast R-CNN. It addresses the limitation of

region proposal generation, a time-consuming step in the previous models.

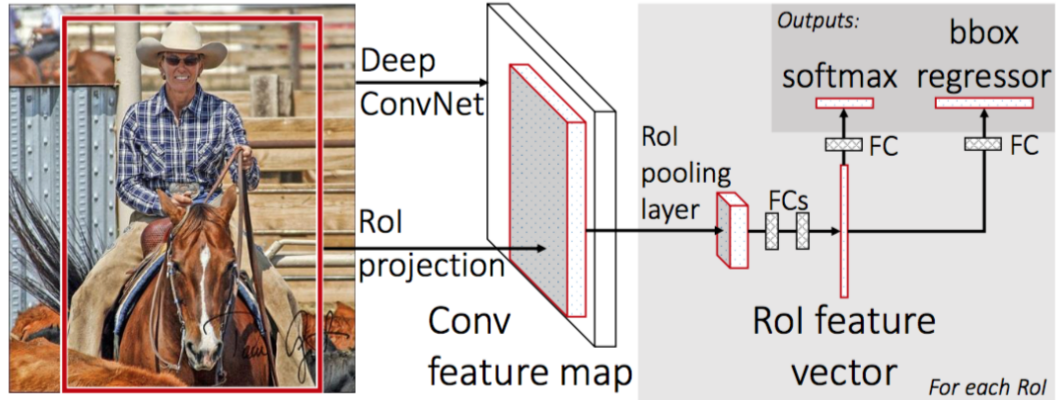


Figure 2.3: A visual representation of Fast R-CNN [3].

Faster R-CNN integrates an RPN within the DNN, facilitating highly efficient operation by utilizing shared features across the detection network. This allows Faster R-CNN to simultaneously predict object proposals and their corresponding bounding boxes. This integration streamlines the process, eliminating the need for selective search or external region proposal methods. Importantly, the RPN incorporates the ROI pooling layer, which uniformly extracts feature vectors of the same size from all proposals within a single image, ensuring robustness in object detection by aligning the proposed regions with the feature map. The RPN efficiently generates region proposals, enhancing the speed and accuracy of object detection. It achieves state-of-the-art results in object detection while significantly reducing computation time, making it highly effective for numerous real-world applications. Fig. 2.4 shows a graphical illustration of Faster R-CNN.

One-stage Object Detectors

While two-stage detectors have demonstrated remarkable success, one-stage object detectors provide an alternative paradigm that simplifies the detection pipeline. In contrast to the two-stage detectors, one-stage detectors integrate region proposal and classification into a single step, aiming to reduce computational complexity and improve real-time performance. This category of detectors is particularly beneficial for real-time applications demanding high speed, such as robotics and video analysis. Examples of

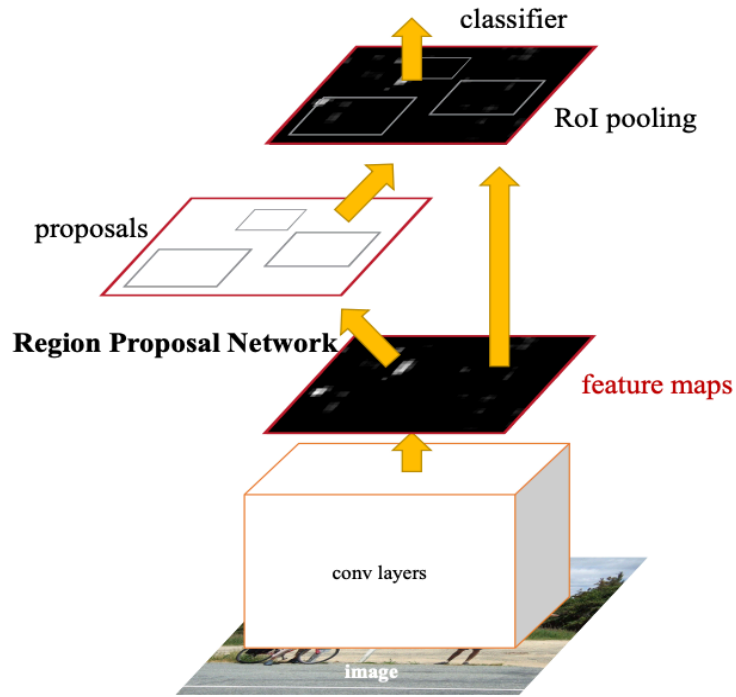


Figure 2.4: A visual representation of Faster R-CNN [4].

detectors in this category include YOLO [5], SSD(Single Shot MultiBox Detector) [6], RetinaNet [37], EfficientNet [38], and DETR(DEtection TRansformers) [7].

YOLO revolutionizes object detection as a one-stage detector that simplifies the task by treating it as a single regression problem. Unlike traditional methods that break the task into multiple steps, YOLO processes the entire image at once using a single neural network, making it highly efficient and suitable for real-time applications. As shown in Fig. 2.5, YOLO divides the image into a grid, simultaneously identifying bounding boxes and class probabilities for each grid cell. This method significantly reduces computation time and allows for high-speed detection, making YOLO ideal for applications such as video analysis, autonomous navigation, and industrial automation, where both speed and accuracy are critical.

SSD is another widely-used one-stage object detection framework known for its balance between speed and accuracy. SSD enhances traditional detection methods by generating a predefined set of bounding boxes with diverse aspect ratios and scales across multiple feature maps at varying resolutions. This multiscale approach enables SSD

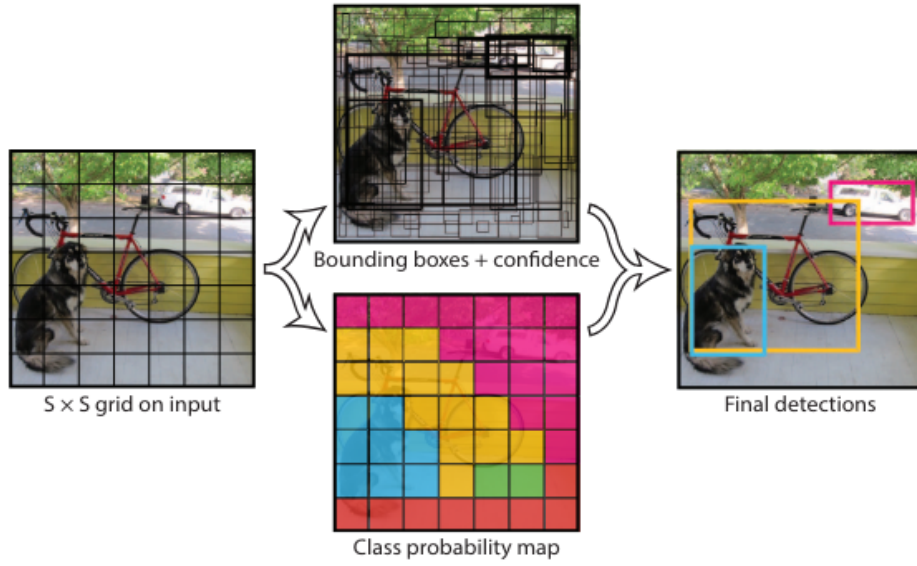


Figure 2.5: A visual representation of YOLO [5].

to effectively detect objects of different sizes. Compared to two-stage detectors, SSD achieves faster detection speeds by integrating feature extraction and bounding box prediction within a single network. Its efficiency and reliability make it ideal for applications such as mobile platforms, real-time video processing, and embedded systems, where computational resources are constrained. The structure of SSD is illustrated in Fig. 2.6.

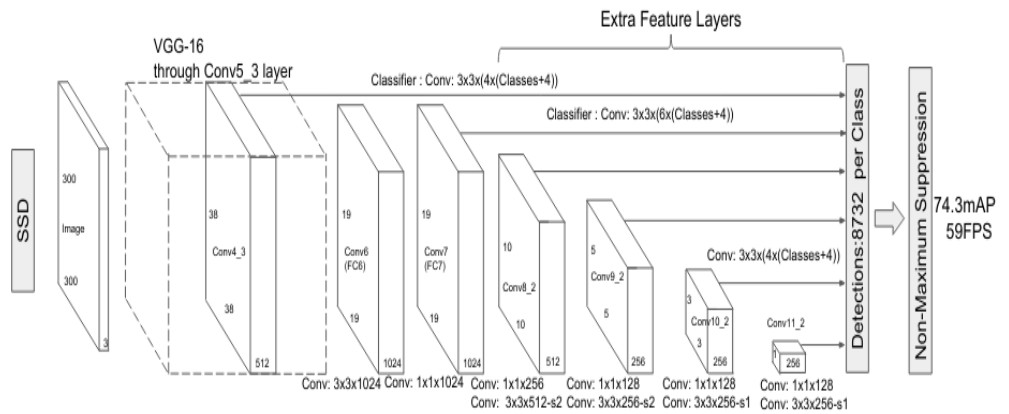


Figure 2.6: Architecture of SSD [6].

DETR introduces a revolutionary approach to object detection by employing trans-

former architectures typically used in natural language processing. As a one-stage detector, DETR eliminates the need for region proposal and postprocessing steps. It treats the entire image as a sequence, directly predicting object bounding boxes and class labels in one step. Leveraging the transformer’s attention mechanism, DETR captures global context and object relationships, thereby enhancing detection performance. By simplifying the pipeline and removing the dependency on anchor boxes, DETR represents a significant shift in object detection methodologies, offering a more straightforward and scalable solution for complex detection tasks in diverse applications such as autonomous driving, surveillance, and medical imaging. Fig. 2.7 shows a visual depiction of DETR.

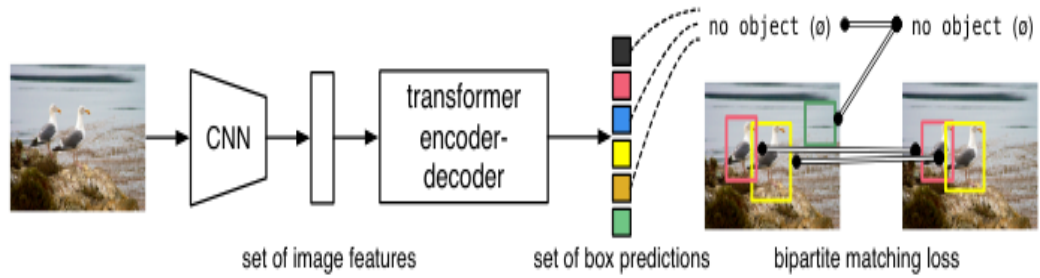


Figure 2.7: A visual depiction of DETR [7].

2.2.2. Instance Segmentation Approaches

As previously discussed, instance segmentation plays a critical role in computer vision. Similar to object detection, instance segmentation identifies and precisely locates distinct objects in an image. However, it goes beyond delineating the exact boundaries of these objects through pixel-level labeling. In contrast to semantic segmentation, which groups pixels based on broad categories without distinguishing between multiple instances of the same class, instance segmentation distinguishes between individual occurrences of objects within the same category. This approach results in an image where every pixel is categorized according to the specific object it belongs to.

Instance segmentation can be approached through two primary methods: detection-

first and segmentation-first [39]. In detection-first methods, the process starts with identifying objects of interest and subsequently segmenting the pixels within the identified regions. On the other hand, segmentation-first methods employ unsupervised techniques where each pixel's label is initially classified at the pixel level, followed by segmentation into individual instance-level objects.

Detection-first Methods

Detection-first methods, a popular approach in instance segmentation, begin with an object detection framework. This framework initially detects objects using bounding boxes. Recent techniques mainly utilize approaches based on R-CNN and their variations. These methods start by generating candidate regions through selective search algorithms, then use CNNs to determine the object's class, thereby completing instance segmentation.

Detection-first methods, widely adopted in instance segmentation, commence with an object detection framework. This framework begins by detecting objects using bounding boxes. Modern techniques often utilize R-CNN-based approaches and their derivatives. Like R-CNN, these methods generate region proposals through selective search algorithms. Subsequently, CNNs are employed to classify the object's category, thereby achieving instance segmentation.

Mask R-CNN extends Faster R-CNN by introducing a parallel branch for predicting object masks alongside bounding box prediction. Unlike RoI pooling in Faster R-CNN, which involves coarse spatial quantization during feature extraction, Mask R-CNN employs the RoI Align method. This method, replacing RoI pooling's harsh quantization with bilinear interpolation [40], ensures precise alignment of feature extraction with the underlying feature map. This alignment preserves spatial information, enhancing pixel-wise segmentation accuracy. Mask R-CNN excels in generating accurate segmentation masks for each identified object, making it highly effective for instance segmentation tasks. It is widely recognized for its role in applications demanding detailed object comprehension and pixel-level precision, solidifying its importance in the field of computer vision. Fig. 2.8 provides a visual overview of Mask R-CNN.

Another network known as **MaskLab**, introduced by Chen et al. [41], is built upon the Faster R-CNN framework. MaskLab, in contrast to Mask R-CNN, incorporates a

semantic segmentation branch aimed at improving the precision of object masks. It also adopts a shared feature pyramid network (FPN) [42] to facilitate concurrent operation of all three branches, thereby optimizing computational efficiency.

Cascade Mask R-CNN [43] extends Mask R-CNN by using a sequence of CNNs to iteratively refine object proposals and segmentation masks. This approach uses a distinct CNN at each cascade stage to enhance accuracy and decrease false positives.

Hybrid Task Cascade(HTC) [44] is an instance segmentation technique rooted in the R-CNN framework. It exploits local and global context cues to boost object detection across diverse sizes and shapes, thereby achieving higher accuracy than the original R-CNN approach.

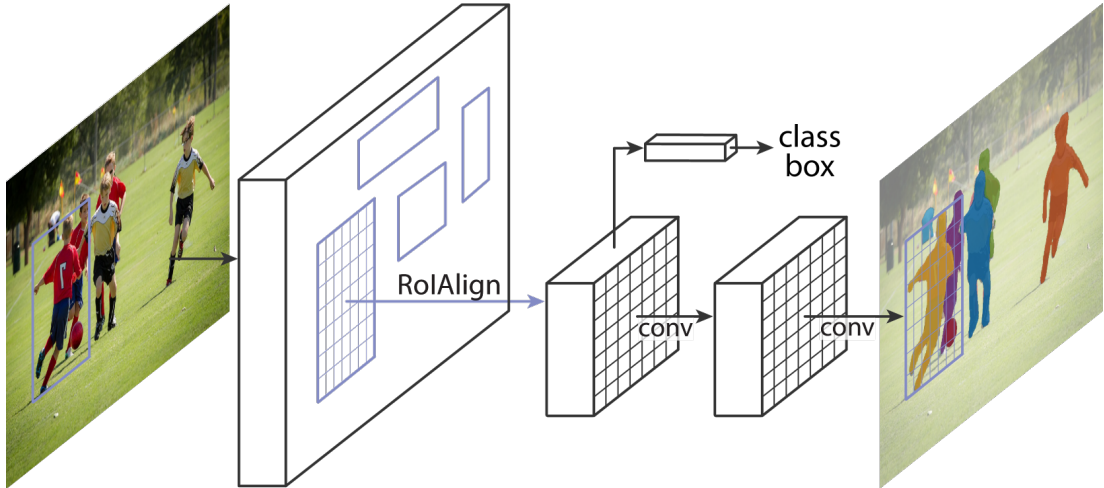


Figure 2.8: A visual representation of Mask R-CNN [8].

Segmentation-first Methods

Segmentation-first methods initiate by conducting semantic segmentation on images to classify each segment into its respective class. An illustration of this is found in the dynamic instantiated network (DIN) [45, 46], which suggests the utilization of distinct subnetworks for semantic segmentation and instance segmentation. This approach dynamically assigns shape and global terms to segmented pixels, addressing challenges posed by nearby instances. By leveraging semantic segmentation, these methods extract shapes and features within each segment to distinguish individual instances. Gao et al. [47] introduce the SSAP (single-shot instance segmentation with affinity pyra-

mid) model, which makes use of affinity pyramids to capture spatial connections between pixels, thereby aiding in instance identification without the need for proposal generation. Within segmentation-first approaches, the accuracy of distinguishing adjacent instances within the same category relies on the segmentation's ability to precisely capture image features.

2.2.3. Object Detection in Remote Sensing

Object detection plays a pivotal role in extracting valuable information from remote sensing imagery. It enables identifying and localizing specific objects or features within vast and complex scenes, fostering applications such as land cover/use mapping, disaster monitoring, precision agriculture, and environmental monitoring.

However, object detection in remote sensing presents unique challenges. The intricate nature of remote sensing data introduces significant variations in scale, objects with arbitrary orientations, weak responses from features, and the presence of tiny objects. Moreover, annotating remote-sensing images demands expertise and is a resource-intensive process.

Early approaches to object detection in remote sensing initially leaned upon traditional computer vision techniques such as template matching [48, 49] and prior knowledge [50, 51], leading to unstable results. Subsequently, machine learning (ML) techniques [21, 52] became prevalent in remote sensing object detection, treating object detection by employing an ML model to search for object proposals and extract features. However, these methods faced limitations in accurately representing objects in challenging scenarios.

Recent developments in DL methodologies have led to the widespread adoption of DL-based approaches in remote sensing object detection. These advancements have significantly enhanced detection performance, primarily attributed to their capacity to extract sophisticated semantic features and resilient object representations. Moreover, these methodologies are deemed particularly suitable for tackling the obstacles introduced by the era of extensive data in remote sensing.

In line with the prominent challenges encountered in remote sensing object detection, existing DL-based remote sensing object detection can be segmented into five cate-

gories: rotated object detection [53], tiny object detection [16], weak object detection [54], multiscale object detection [55], and object detection with limited supervision [19]. A comprehensive examination of DL-based object detection in remote sensing can be accessed in [56].

2.3. Domain Generalization

ML models have achieved remarkable success in a variety of fields, including computer vision and natural language processing. These models are designed to learn general and predictive patterns from training data and then utilize this knowledge to make predictions on previously unseen test data. The standard procedure involves the training of ML models on the premise that the data for training and testing adhere to an identical and independent distribution, often denoted as i.i.d., notwithstanding the fact that this assumption may not hold true in practical settings. When there is a mismatch between the distributions of the training and test data, known as domain shift, the performance of ML models often declines. Domain shift can arise due to various factors such as differences in lighting conditions, viewpoints changes, or sensor types.

One straightforward solution to mitigate the domain shift problem is to collect more data to improve the quality and diversity of the training set. However, acquiring additional data is often challenging and time-consuming, and it is unrealistic to expect to cover all possible deployment scenarios.

Many generalization-related research topics address the domain shift problem, such as domain adaptation, domain generalization, transfer learning, meta-learning, etc. Among these, domain generalization has garnered significant attention recently. As depicted in Fig. 2.9, domain generalization addresses the domain shift problem by learning a model from one or several different but related domains that can perform well on unseen testing domains. For detailed surveys on domain generalization, see [57] and [58].

2.3.1. Problem Definition

Definition 1 (Domain). Let A denote the input space and B the output space. A *domain* is constituted by data sampled from the joint distribution $S = \{(a_i, b_i)\}_{i=1}^n \sim P(A, B)$ on $A \times B$, where $a \in A \subset \mathbb{R}^d$, $b \in B \subset \mathbb{R}$, and n is the number of samples.

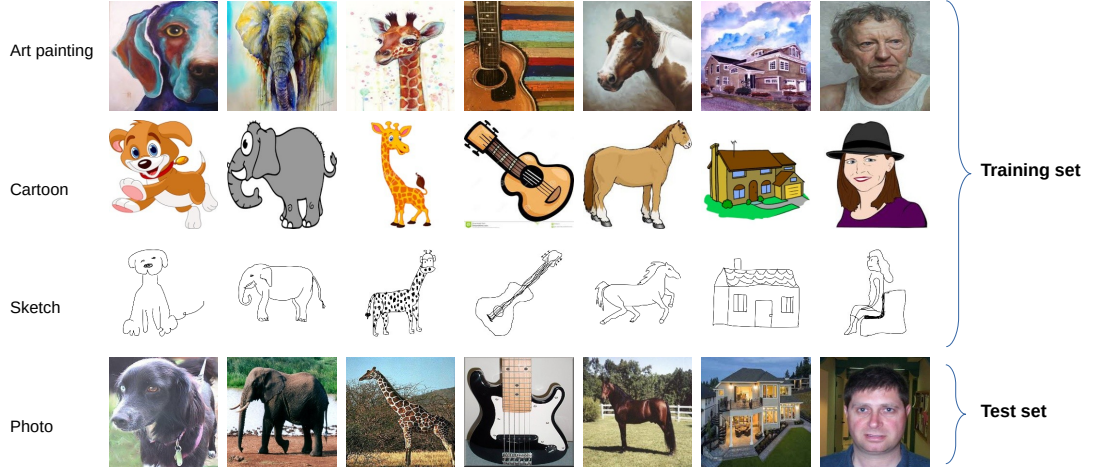


Figure 2.9: Samples from the PACS [9] dataset for domain generalization. The training set comprises art paintings, cartoons, and sketch images, while the test set comprises photographic images. DG aims to learn from the training set to perform well on the unseen test set.

Definition 2 (Domain generalization). Within the domain generalization, we have access to K training (source) domains, denoted as $S_{train} = \{S^i \mid i = 1, \dots, K\}$ where $S^i = \{(x_j^i, y_j^i)\}_{j=1}^{n^i}$ represents i -th domain. Note that $P^{(i)}(X, Y) \neq P^{(j)}(X, Y)$ with $i \neq j$ and $i, j \in \{1, \dots, K\}$. The primary goal of DG is to develop a predictive function $h : X \rightarrow Y$ that exhibits strong generalization performance across various domains. This function is trained using data from K source domains to reduce the prediction error when applied to an unseen test (target) domain S_{test} . Note that S_{test} is not accessible during training and $P^{(test)}(X, Y) \neq P^{(i)}(X, Y)$ for $i \in \{1, \dots, K\}$.

2.3.2. Related Research Areas

Multi-task Learning(MTL). The main objective of MTL [59] is to encourage a model to learn and perform multiple related tasks simultaneously [60, 61]. In contrast to DG, which focuses on generalizing a model to unseen data distributions ($P_s(X, Y) \neq P_t(X, Y)$), MTL enhances the model's ability to perform well on a specific collection of tasks that it has been trained on ($Y_s = Y_t$). Despite the differences in their problem setups, some domain generalization methods leverage the principles of MTL [62, 63].

Transfer learning. Transfer learning (TL) [64] entails training a model on one domain or task to gain knowledge that can enhance its performance on a different but related target domain or task. A common example of TL is fine-tuning, where DNNs are initially pre-trained on extensive datasets like ImageNet [65] and subsequently fine-tuned for a specific downstream task. While TL necessitates access to target data for adapting the model to new tasks, DG operates under the assumption that no target data is available, concentrating instead on generalizing across domains. It is important to note that both TL and DG involve a shift in the target distribution compared to the source distribution.

Domain Adaptation. Domain adaptation [66, 67] has gained significant popularity in recent years and is closely related to DG, with substantial research dedicated to it [68–70]. Both DA and DG aim to address the issue of domain shift. However, the crucial difference between them lies in the fact that DA utilizes target domain data during the training phase, whereas DG does not, rendering DG a more challenging yet pragmatic approach compared to DA.

2.3.3. Domain Generalization Methods

This section classifies existing DG methods according to their methodologies: data manipulation, representation learning, and learning strategy.

Data manipulation. The importance of the training data’s quantity and variety cannot be overstated when considering the ability of ML models to generalize effectively. A straightforward approach to enhancing training data diversity is through data manipulation, which simultaneously increases data quantity. The basic idea in data-manipulation-based DG methods is to expand the initial (a, b) pairs by introducing new $(T(a), b)$ pairs, where $T(\cdot)$ represents a transformation, usually preserving the labels. Based on the difference in transformation function, domain manipulation-based DG techniques are categorized as either *data augmentation* or *data generation*.

Data augmentation is a technique to enhance the generalization capacity of ML models and avoid over-fitting. Common augmentation techniques used in supervised learning are rotation, flipping, cropping, scaling, adding noise, etc. They can also be applied for DG without exception to improve generalization [71]. In addition to the typical aug-

mentation methods applied in DG, other notable approaches include domain randomization [72–75]—introducing variability by randomizing environmental factors during training—and adversarial domain augmentation [76], [77], [78], aiming to optimize generalization by enhancing data diversity while ensuring reliability through a strategic augmentation process.

Representation learning. Representation learning involves extracting meaningful features from data. In conventional supervised learning, the goal is to learn a representation that minimizes the error in the training data. However, in DG, the target domain remains unknown during training. Therefore, the representation needs to exhibit robustness to variations in the data distribution and capture domain-invariant features.

More formally, the optimization problem can be formulated as:

$$\min_{f,g} \mathbb{E}_{x,y} [\mathcal{L}(f(g(x), y)) + \lambda \mathcal{L}_{\text{reg}}] \quad (2.1)$$

Here f , g , λ , and $\mathcal{L}_{\text{reg}}$ denote the prediction function, representation function, tradeoff parameter, and regularization term, respectively. The most popular domain generalization methods fall into this category, specifically classified as *Feature disentanglement* and *Domain-invariant representation learning*.

The primary goal of domain-invariant representation learning-based DG methods is to acquire a representation that remains invariant to domain shift among source domains, focusing on capturing the underlying domain-invariant information. The straightforward idea behind these methods is that features resilient to variations in source domains will also be robust against shifts in any unseen target domains. This category of methods can be further classified into kernel-based methods, explicit feature alignment, and domain adversarial learning.

Kernel-based methods [79–81] leverage a kernel function as the representation function, implicitly transforming original data into a higher-dimensional space where domain-invariant features can be effectively extracted.

DG methods based on explicit feature alignment aim to reduce discrepancies between source domains by explicitly aligning their feature distributions. This enables the learning of domain-invariant representations. Various metrics are employed to quantify

distribution differences and guide alignment. Common metrics include minimizing moments [82, 83], minimizing contrastive loss [84, 85], minimizing KL divergence [86, 87], and minimizing maximum mean discrepancy [88, 89].

Domain adversarial learning is a commonly used technique in DG. Its primary objective is to train a feature extractor and a domain discriminator adversarially to learn features that are independent of specific domains. The concept is to align features so that the domain discriminator cannot differentiate them. This discriminator can be designed as a multi-class classifier [90, 91] or as a binary classifier for each domain [92]. Training involves alternating between updating the feature extractor and the domain discriminator. Alternatively, a simplified approach involves a single-step update using a gradient reversal layer (GRL)[93]. Domain adversarial learning techniques in DG can also be integrated with other methods [86, 94].

In contrast to methods that enforce domain-agnostic representations, feature disentanglement, another representation learning-based DG method, aims to decompose features into domain-invariant and domain-specific components. Generative learning and multi-component analysis are two types of these methods, categorized based on their network architecture and implementation strategies.

Generative learning methods [95, 96], on the other hand, utilize generative models to learn a latent representation that encapsulates the domain-invariant features. These models are trained to generate data samples from the latent representation, ensuring that the generated samples retain the data’s essential characteristics while invariant to domain-specific variations.

Multi-component analysis methods [97, 98] employ multiple neural networks to extract domain-invariant and domain-specific features separately. One network focuses on capturing the underlying concepts shared across domains, while the other concentrates on capturing the distinctive characteristics of each domain.

Learning strategy. Besides representation learning and data manipulation methods, several learning strategies have been developed to promote domain generalization. These techniques aim to enhance the robustness of models against domain shifts by utilizing tailored algorithms or training procedures. This class of methods can be broadly categorized into three groups: self-supervised learning, ensemble learning, and meta-learning.

Self-supervised learning techniques pre-train models on large amounts of unlabeled data using pretext tasks that do not rely on task-specific annotations. These tasks encourage the model to learn generalizable representations that can be effective in different domains. A notable advantage of self-supervised learning is its applicability in single-source and multi-source scenarios, eliminating the need for domain labels. For instance, [99] exemplifies a study demonstrating the application of self-supervised learning in domain generalization.

Ensemble learning combines multiple models trained on different domain-specific representations or with diverse training strategies. This diversity can enhance the model’s capture of shared information across domains. Various ensemble techniques exist, such as domain-specific neural networks [100, 101], where each model specializes in a particular domain. The final prediction is obtained by combining the predictions of individual experts weighted by their confidence. Another technique is weight-averaging [102], where the predictions of multiple models trained on different source domains are averaged, resulting in a less sensitive model to domain-specific biases.

Meta-learning aims to learn ”how to learn” by training a model to adapt quickly to new domains encountered during testing. This objective is accomplished by partitioning the data from the source domain into meta-train and meta-test sets to replicate domain shift. Within this category, methods can be more specifically categorized into model-based [103] approaches, optimization-based techniques [104], and metric-based methodologies [105].

2.4. Domain Generalization for Remote Sensing

While domain generalization in remote sensing has received limited attention, several researchers have made noteworthy contributions. Zhu et al. [106] proposed SCSN, a content-focused network with a dedicated loss function for improved cross-scene generalization. Luo et al. [107] introduced WHU-Mix, an extensive building dataset, and BSM, an easily integrable domain generalization method for enhanced building extraction. Zhang et al. [108] presented LDGnet, a language-aware network that leverages text descriptions to improve hyperspectral image classification. Gao et al. [109] developed LiCa, which aligns material distributions across scenes for robust pixel-wise clas-

sification. Zheng et al. [110] proposed MMD-DRCN, which learns multi-source representations and refines features through reconstruction for generalizable oil palm tree detection. Zhao et al. [111] introduced DAD3GM, a model combining dual-attention and deep discriminative feature learning for state-of-the-art hyperspectral image classification accuracy without labeled target samples. Liang et al. [112] introduced CCDR, a method for single-domain generalization that improves remote sensing image segmentation through domain randomization, outperforming other SDG methods in this field. Wang et al. [113] proposed a technique using nonlinear sample generation for hyperspectral cross-scene classification, with contrastive learning ensuring the quality of the generated samples. Lizuka et al.[114] proposed the FOSMix method, which enhances domain generalization in remote sensing segmentation, with its components—FM, OM, and regularization—ensuring superior segmentation performance. Qi et al. [115] introduced MDGTnet for image classification, including loss function variations such as intra-DUEB, inter-DCEB, and a classifier.

3. DOMAIN GENERALIZED OBJECT DETECTION FOR REMOTE SENSING

3.1. Introduction

In recent years, advancements in sensor technologies have significantly improved the quality and quantity of high-resolution remote sensing images. These technological strides have not only enhanced our ability to capture detailed images of Earth’s surface but have also created new opportunities for automated analysis and interpretation of these images. The ability to accurately detect and classify objects within these images – such as airplanes, buildings, and ships – has become increasingly essential. Such capabilities are particularly crucial in a variety of applications, including environmental monitoring, hazard detection, urban planning, and change detection [116, 117], where precision and reliability are paramount.

However, despite these advancements, object detection within remote sensing images remains a formidable challenge in both remote sensing and computer vision. One of the primary difficulties stems from the phenomenon of domain shift, where the performance of data-driven models deteriorated due to discrepancies between the training and testing data distributions. This issue, illustrated in Figure 3.1, is common in real-world scenarios and can significantly impede the generalization capabilities of models. Addressing this challenge often involves increasing the amount and diversity of training data, which, while helpful, can be both costly and time-consuming. Moreover, this approach does not guarantee comprehensive coverage of all potential deployment scenarios and variations of objects.

Two widely adopted techniques to mitigate the effects of domain shift are DA and DG. Both methods aim to align the data distributions across source and target domains, thereby improving the model’s generalization performance. However, these techniques differ in their approaches: DA requires access to target domain data during training, whereas DG does not. Consequently, DG is often more demanding but also more reflective of real-world conditions, where the model must perform well on unseen data from different domains.

Despite the importance of this issue, research on domain generalization, particularly in the context of remote sensing, has been limited. While there have been studies on domain generalization in remote sensing images for various computer vision tasks [106–111], there is a noticeable gap in research specifically focused on domain generalization for remote sensing object detection. This chapter aims to fill this gap by proposing a novel technique for domain generalization that targets building detection in remote sensing images. The proposed approach is designed to effectively handle domain shift at both the image and instance levels. Additionally, it incorporates consistency regularization to enhance domain invariance within the model, thereby improving its robustness across different domains. To the best of our knowledge, this study represents the first attempt to explore domain generalization for remote sensing images in the specific context of object detection.



Figure 3.1: Buildings from different cities (DFC23 dataset); from top-left to bottom-right: Addis Ababa, New Delhi, Barcelona, Darwin (bottom-left), Brasilia, and New York.

3.2. Preliminaries

This section begins with an outline of the foundational mathematical concepts underpinning the Faster R-CNN model [4], which serves as the basis for the proposed DG method. Subsequently, it delves into the essential components of the proposed DG method designed for object detection.

Faster R-CNN. As outlined in Chapter 2, Faster R-CNN is a two-stage object detector comprising three main components: shared convolutional layers at the base, a Region Proposal Network (RPN), and an ROI-based classifier. The loss function for training incorporates both the classifier loss for ROI and the RPN loss:

$$\mathcal{L}_{det} = \mathcal{L}_{roi} + \mathcal{L}_{rpn} \quad (3.1)$$

The training loss for both the ROI and RPN classifiers comprises two main components: one evaluates the accuracy of classification by assessing predicted probabilities, and the other improves bounding box localization through regression.

Object detection. Object detection involves modeling the posterior probability $P(Y, O|x)$, where x denotes the input image. Here $O \in \mathbb{R}^4$ represents the bounding box coordinates of an object and $Y \in \{1, \dots, M\}$ indicates the category of the corresponding object, where M denotes the total number of possible classes.

Domain generalization. Consider the collective distribution of training data utilized in object detection, represented as $P_D(Y, O, x)$, where D signifies the image domain. When there is a domain shift, the distributions $P_{D_i}(Y, O, x)$ and $P_{D_j}(Y, O, x)$ differ, with $i \neq j$ and $i, j \in 1, \dots, N$, with N being the number of training domains. The DG approach aims to train an object detection model that can extract features invariant to domain variations across these N source domains, allowing it to perform effectively on previously unseen target domain(s).

Image-level generalization. Utilizing Bayes' formula, the joint distribution can be decomposed in the following manner:

$$P(Y, O, x) = P(Y, O|x)P(x) \quad (3.2)$$

When considering covariate shift in object detection, the conditional probability $P(Y, O|x)$ remains constant across different domains. Nevertheless, the disparity between domains is due to variations in the marginal distribution $P(x)$. Simply put, the detector should yield consistent results across domains, ensuring that detection outcomes remain unchanged regardless of the image’s domain. Within the Faster R-CNN framework, the image representation x is synonymous with the resultant feature map stemming from the initial convolutional layers. Consequently, in addressing domain shift, it is imperative to uphold similarity in the distribution of these image representations across diverse domains.

Instance-level Generalization. Conversely, the joint distribution can also be decomposed in the following manner:

$$P(Y, O, x) = P(Y|O, x)P(O, x) \quad (3.3)$$

Under the covariate shift, the conditional probability $P(Y|O, x)$ remains consistent across diverse domains, resulting in a change within the domain distribution in the marginal $P(O, x)$. This indicates that there is a preservation of semantic coherence across domains: the category labels of an object should remain consistent if it is present in the same image region, regardless of the domain. Therefore, we aim to maintain consistency in the distribution of instance representations across domains. The instance representation (O, x) specifically refers to the characteristics extracted from the image region.

3.3. Proposed Method

Image-level Domain Generalization: A method is presented to address domain shift specifically at the image level. Within Faster R-CNN with FPN, the backbone network is responsible for producing five distinct levels of feature maps, as depicted by the green rectangle in Figure 3.2. The initial feature map, serving as the image representation, undergoes further processing through convolution and fully connected layers to enable domain classification at the image level.

Let I_n denote the n^{th} of N training images, D_n its corresponding domain label from $|D|$ possible domains, and p_n the $|D|$ -length vector of predicted probabilities of the

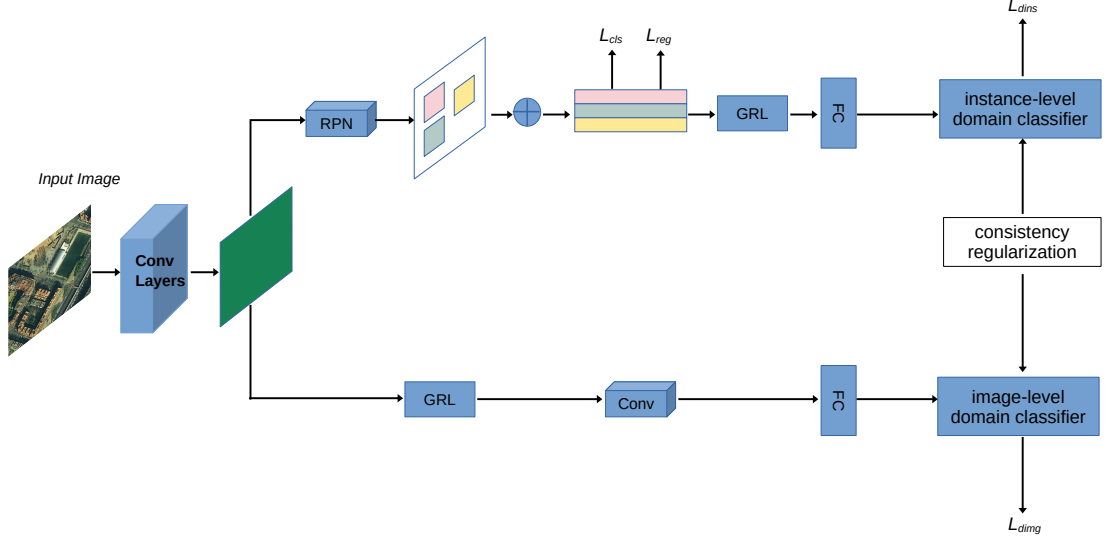


Figure 3.2: An outline of the suggested Domain Generalized Faster R-CNN network, where RPN stands for region proposal network, Conv for convolutional layer, FC for fully connected layer, and GRL for gradient reversal layer.

domain of the image. By employing cross-entropy loss, it is possible to construct the loss function for domain classification at the image level in the following manner:

$$\mathcal{L}_{dimg} = - \sum_{n=1}^N \mathbf{1}_{D_n} \cdot \log(p_n), \quad (3.4)$$

The function $\mathbf{1}$ denotes the one-hot vector indicator.

Aligning image-level representations is generally effective in minimizing the impact of global image differences, such as scale, illumination, and style, thereby reducing the disparity between images.

Instance-level Domain Generalization: ROI-based feature vectors are processed through fully connected layers for instance-level domain classification to reduce domain shift at the instance level.

Let $p_{n,m}$ represent the predicted domain of the m -th region proposal in the n -th image. Let N and M denote the number of images and objects in the training batch, respectively, and let D_n represent the domain label of the n -th image. The instance-level domain classification loss, using cross-entropy loss, can be formulated as:

$$\mathcal{L}_{dins} = - \sum_{n=1}^N \sum_{m=1}^M \mathbf{1}_{D_n} \cdot \log(p_{n,m}). \quad (3.5)$$

where $\mathbf{1}$ is the one-hot vector indicator function. Aligning the instance-level representations can mitigate differences between local instances, such as variations in object appearance, size, viewpoint, and so on.

Consistency Regularization:

Furthermore, a consistency regularization method is employed to ensure coherence between the domain classifications at the image and instance levels. Specifically, the prediction from the image-level domain classifier must match the prediction from the instance-level domain classifier for each instance within the image.

Let p_n and $p_{n,m}$ denote the predicted domain of n -th image and m -th region proposal in the n -th image, respectively. By using mean square error, the consistency loss can be written as:

$$\mathcal{L}_{cons} = \frac{1}{NM} \sum_{n=1}^N \sum_{m=1}^M (p_n - p_{n,m})^2 \quad (3.6)$$

Network Architecture: An outline of the proposed network is depicted in Figure 3.2. To train the network adversarially, a GRL [93] is used. The features are passed through the GRL before being fed into the domain classifier network. When the features pass through the GRL layer, the gradient sign is reversed to optimize the network. The sum of the Faster R-CNN loss and the losses associated with domain generalization determines the final training loss. This loss function is expressed in the following equation:

$$L = L_{det} + \alpha L_{img} + \beta L_{ins} + \gamma L_{cst} \quad (3.7)$$

where α , β , and γ are empirically set trade-off parameters for image-level, instance-level, and consistency regularization losses, respectively.

3.4. Experiments

3.4.1. Dataset

In the context of the IEEE DFC 2023 dataset [118], a large-scale remote sensing image dataset has been provided for detecting building roof types. The dataset includes images from the SuperView-1 and Gaofen-2 satellites, which offer spatial resolutions of 0.5m and 0.8m, respectively. Each image has dimensions of 512×512 pixels and was collected from seventeen different cities across six continents, labeled according

to twelve distinct categories of roof types. Each city represents a unique domain in the dataset. Tables 3.1 and 3.2 present the distribution of images and instances across continents and cities, as well as the distribution of roof type categories.

Table 3.1: Distribution of images and instances across various cities on different continents.

Continent	City	# of images	# of instances
Africa	AddisAbaba	57	4871
	NewDelhi	135	5929
Asia	Suzhou	437	8961
	Tokyo	475	15617
Europe	Barcelona	227	20160
	Berlin	566	22232
	Copenhagen	314	16372
	Portsmouth	155	19362
Austria	Darwin	230	9985
	Sydney	36	951
S. America	Brasilia	248	9864
	Rio	97	8836
	SaoPaulo	116	12384
	SaoLuis	30	9645
N. America	NewYork	198	13528
	Jacksonville	291	6176
	SanDiego	108	9390
Total		3720	194263

3.4.2. Experiment Setup

Table 3.2 demonstrates a significant imbalance in the distribution of roof types within the dataset. Two strategies are implemented to address this imbalance and prioritize domain generalization: the two-roof types and building-only settings.

Table 3.2: Distribution of roof type categories in the IEEE DFC 2023 dataset, showing the number of instances and their respective percentages.

No	Roof type category	# of instances	Percent
1	Flat	109640	56,44
2	Gable	57426	29,56
3	Gambrel	5087	2,62
4	Row	87326	0,49
5	Multiple eaves	738	0,38
6	Hipped v1	4320	2,22
7	Hipped v2	10220	5,26
8	Mansard	510	0,26
9	Pyramid	1520	0,78
10	Arched	833	0,43
11	Revolved	154	0,08
12	Other	2872	1,48
Total		194263	100

Two-roof types Setting: This approach retains only the two most prevalent roof types, Flat and Gable, while excluding all other categories. Three cities from Asia are selected for model validation, and four cities from South America are designated for testing. The training set comprises images from the remaining cities.

Building-Only Setting: In this approach, all 12 distinct roof types are consolidated into a single class, representing buildings as a whole. For this setting, a three-fold cross-validation experiment was conducted to validate the cross-domain performance of the model. The first fold followed the same setup as in the two-roof types setting, with cities from Asia for validation and cities from South America for testing. In the second fold, cities from Asia were used for validation, and cities from North America for testing. In the third fold, European cities were used for validation, and Asian cities were used for testing.

Three experiments—a lower baseline, an upper baseline, and domain generalization—are conducted to compare and evaluate the proposed domain generalization method.

Lower baseline: This experiment employs the standard Faster R-CNN object detection method without incorporating any domain generalization techniques. The optimal model is selected based on performance metrics from 80% of the validation set, and its effectiveness is assessed using 80% of the test set.

Oracle: This experiment aims to establish the upper limit of performance. Unlike the lower baseline, 20% of the validation dataset and 20% of the test dataset are included in the training data.

Domain generalization: This experiment employs the proposed domain generalization technique. The neural network undergoes training using the training dataset, validation using 80% of the validation dataset, and performance evaluation using 80% of the test dataset. Moreover, findings are presented for various configurations involving image-level and instance-level domain alignments alongside consistency regularizations.

Training: Based on empirical observations, the tradeoff parameters were configured as $\alpha = 0.1$, $\beta = 0.1$, and $\gamma = 0.1$. Early stopping was employed with a patience of 10 epochs to prevent overfitting. Stochastic Gradient Descent (SGD) was chosen as the optimizer with specific configurations: a weight decay of 0.0001, momentum of 0.9, learning rate of 0.01, and a batch size of 16. The experiments were conducted using the Torchvision Faster R-CNN library of PyTorch framework, utilizing an NVIDIA GeForce GTX 1080 Ti equipped with 12 GB of GPU memory.

3.4.3. Results

The proposed method was assessed using the mean average precision (mAP) metric, measured both at a single threshold (0.5: mAP@50) and across multiple thresholds (0.5-0.95: mAP). The results, presented in Table 3.3 and Table 3.4, demonstrate consistent enhancements in performance over the baseline (Faster R-CNN) with both instance-level and image-level domain generalization techniques. Specifically, in the buildings-only setting, instance-level generalizations outperformed image-level generalizations, whereas image-level generalization achieved greater performance improvements in the two-roof types setting. The highest performance gains were observed when combining instance-level and image-level domain generalization techniques. Furthermore, consis-

Table 3.3: mAP and mAP at a 0.5 IoU threshold results on the IEEE DFC 2023 dataset for the two predominant categories: Flat and Gable. All other roof types are excluded from the dataset.

Method	image	instance	consistency	mAP	mAP@50
Faster R-CNN (lower baseline)				4.6	9.1
Faster R-CNN (oracle)				12.3	21.9
DGFRCNN-RS(proposed)	✓			5.9	11.7
		✓		5.2	10.5
	✓	✓		6.8	13.3
	✓	✓	✓	7.1	14.2

tency regularization further bolstered performance in both scenarios. More specifically;

Two-roof types setting

In the two roof types setting, the lower baseline achieved an mAP of 4.6 and a mAP@50 of 9.1. The proposed domain generalization method significantly outperformed the baseline, with image-level generalization alone yielding an mAP of 5.9 and a mAP@50 of 11.7. Instance-level generalization reached an mAP of 5.2 and a mAP@50 of 10.5. The combined approach resulted in an mAP of 6.8 and a mAP@50 of 13.3. Finally, the addition of consistency regularization improved the results further, achieving an mAP of 7.1 and a mAP@50 of 14.2.

Building-only setting

The results of the three-fold cross-validation experiments in the building-only setting are presented in Table 3.5. The performance of the proposed domain generalization method (DGFRCNN-RS) is compared against the lower baseline (standard Faster R-CNN) and the upper baseline (oracle). The lower baseline, which uses the standard Faster R-CNN object detection method without any domain generalization techniques, shows an average mAP of 14.57 and an average mAP@50 of 30.7. These results indicate the basic performance level of the model without any enhancements for domain generalization. The results across the three folds show some variability, which is expected due to differences in the validation and testing domains.

Table 3.4: Cross validation experiment results for building detection using different folds with the proposed DGFR CNN-RS model. Each fold used images from different continents for training, validation, and testing. The performance metrics reported include mAP and mAP at a 0.5 IoU threshold (mAP@50).

	image	instance	consistency	mAP	mAP@50
Faster R-CNN (lower baseline)					
Fold 1				10.6	22.3
Fold 2				18.3	36.1
Fold 3				14.8	33.7
Average				14.57 ± 3.86	30.7 ± 7.37
Faster R-CNN (oracle)					
Fold 1				18.9	35.8
Fold 2				22.9	45.6
Fold 3				20.4	40.5
Average				20.73 ± 2.02	40.63 ± 4.9
DGFR CNN-RS (proposed)					
Fold 1				10.8	22.7
Fold 2	✓			18.6	36.9
Fold 3				15.6	34.6
Average				15 ± 3.93	31.4 ± 7.4
DGFR CNN-RS (proposed)					
Fold 1				11.1	22.9
Fold 2		✓		18.8	37.7
Fold 3				15.2	29.9
Average				15.03 ± 3.85	30.17 ± 7.4
DGFR CNN-RS (proposed)					
Fold 1				11.2	23.1
Fold 2	✓	✓		18.9	37.1
Fold 3				15.8	30.8
Average				15.3 ± 3.87	30.33 ± 7.01
DGFR CNN-RS (proposed)					
Fold 1				11.5	23.4
Fold 2	✓	✓	✓	19.2	38.3
Fold 3				16.3	32.6
Average				15.67 ± 3.89	31.43 ± 7.52

The Oracle experiment, which includes a portion of the validation and test data in the training set, sets an upper-performance limit with an average mAP of 20.73 and an average mAP@50 of 40.63. This significant improvement over the lower baseline highlights the potential performance gains if the model had prior knowledge of the target domains.

The proposed DGFCNN-RS method consistently improves over the lower baseline, incorporating image-level, instance-level, and consistency regularization techniques. The average mAP and mAP@50 across the three folds for the DGFCNN-RS are 15.67 and 31.43, respectively. This demonstrates the effectiveness of the domain generalization approach in enhancing model performance when dealing with diverse and unseen domains.

Overall, these findings underscore the efficacy of the proposed domain generalization approach, particularly when combining instance-level and image-level techniques, which is further enhanced by consistency regularization.

3.4.4. Ablation Study

The ablation study presented in Table 3.5 evaluates the impact of different weights (α, β, γ) on the model’s performance metrics, specifically mAP and mAP at a 0.5 IoU threshold (mAP@50). The results indicate that the highest mAP (11.5) and mAP@50 (23.4) is achieved when all three weights (α, β , and γ) are set to 0.1. This balanced setting appears to optimize the contributions of image-level, instance-level, and consistency regularization losses, resulting in improved overall performance. Conversely, significantly altering any single weight (e.g., $\alpha = 0.4$ or $\beta = 0.4$) tends to degrade performance, suggesting that a balanced approach is most effective for this model and dataset.

Given these findings, experiments with the validation set are suggested to identify the optimal values for these variables. By systematically testing different weight combinations on the validation set, one can tailor the model’s performance to better suit their specific data and application requirements.

Table 3.5: Results of the ablation study for weights α , β , and γ in Equation 3.7, showing the impact on mean average precision (mAP) and mAP at a 0.5 IoU threshold (mAP@50) for the proposed DGFRNN-RS model.

α	β	γ	mAP	mAP@50
0.025	0.1	0.1	11.2	22.8
0.1	0.1	0.1	11.5	23.4
0.4	0.1	0.1	11.4	23
0.1	0.025	0.1	10.5	23
0.1	0.4	0.1	10.4	22.1
0.1	0.1	0.025	11.2	22.9
0.1	0.1	0.4	10.3	22.1

3.4.5. Conclusion

A domain generalization method is introduced in this chapter to tackle domain shift at both the instance and image levels in remote sensing image object detection. Additionally, a consistency regularization technique is proposed to enhance the domain-invariance of the generalization model. The effectiveness of the proposed approach is evaluated using the IEEE DFC 2023 dataset in two different settings: one where all 12 different roof types are treated as buildings and another where only the two predominant categories, Flat and Gable, are considered. In both cases, the findings indicate that aligning source and target domains at both instance and image levels significantly boosts performance. Furthermore, even greater performance improvements result from combining both alignment types. These results are further enhanced by consistency regularization. Various domain alignment approaches will be explored, and the proposed method will be evaluated on different datasets in future work.

4. A DOMAIN GENERALIZED MASK R-CNN FOR BUILDING INSTANCE SEGMENTATION

4.1. Introduction

Thanks to advances in sensor technologies, high spatial resolution remote sensing images have proliferated and have become more accessible, paving the way for more complex processing such as object detection and instance segmentation. Object detection means locating and classifying objects using bounding boxes, while instance segmentation is a step further and generates pixel-level masks that accurately outline the shape of each object instance. The ability to segment objects of interest at pixel-level precision has direct applications in urban planning as building [119] and vehicle detection [120], in maritime traffic monitoring ship detection [121].



Figure 4.1: Buildings from different cities (DFC23 dataset); from top-left to bottom-right: Addis Ababa, New Delhi, Barcelona, Darwin (bottom-left), Brasilia, and New York.

Instance segmentation approaches are generally categorized as either top-down, bottom-up or direct [122]. While top-down approaches such as Mask R-CNN [8] first detect then segment, bottom-up strategies, first segment and then cluster into instances [123]. Finally, direct methods conduct pixel aggregation and prediction concurrently [124]. Top-down methods – with Mask R-CNN variants in particular – have gained significant popularity due to superior performance against complex scenes and object distributions. In the particular context of building instance segmentation, boundary optimization with context-aware modules has been recently investigated [119]. Moreover, instance segmentation methods have also been adapted to hyperspectral remote sensing images [125] in combination with spectral-spatial FPNs. Composite dual backbones have further been explored [126].

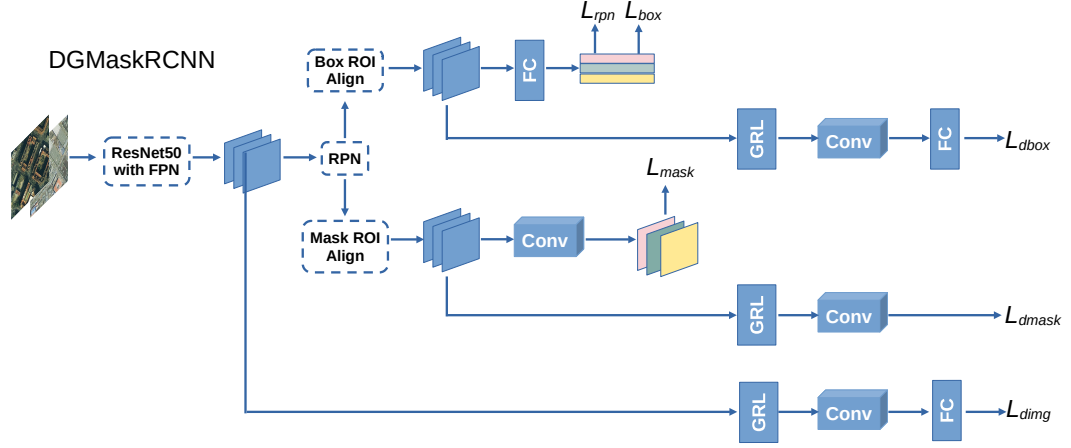


Figure 4.2: An outline of the proposed Domain Generalized Mask R-CNN model where FPN, RPN, Conv, FC, and GRL denote feature pyramid network, region proposal network, convolutional layer, fully connected layer, and gradient reversal layer, respectively.

Nevertheless, real-world applications frequently encounter a challenge known as domain shift, where the performance of a data-driven model deteriorates due to disparities between the training (source) and testing (target) data distributions (Fig. 4.1). Expanding the volume and variety of training data is a possible approach to mitigate this issue, but it can be expensive and time-consuming. Furthermore, it does not guarantee that every possible deployment scenario and object variation will be adequately represented.

The issue of domain shift has been tackled in cutting-edge research through diverse forms of TL, such as knowledge distillation [127], and predominantly through DA [126], which modifies parameters learned from a training domain to work better in the new test domain, given some example data from it. In contrast, DG refers to the setting where no such example test domain data is accessible and aims to learn a single model from one or several different but related domains that can perform well on unseen test domains. However, DG has received limited attention in the context of remote sensing [106, 107, 110, 128].

This study takes the first step toward a domain-generalized model for building instance extraction from remote sensing images. Following a top-down instance segmentation paradigm, a Mask R-CNN model is equipped with image-level, instance-level, and pixel-level domain adversarial modules to promote extracting domain-invariant features. The proposed approach is validated through the 2023 IEEE DFC dataset [10], where preliminary results obtained in the context of cross-continent generalization show superior performance over the baseline Mask R-CNN.

4.2. Domain generalized Mask R-CNN

Mask R-CNN [8] is an extension of Faster R-CNN, for instance segmentation. Mask R-CNN is a two-stage detector that employs ResNet-50 [129] as a backbone and an FPN [42] as the neck to extract multi-scale features. An RPN proposes candidate object regions, refined by an ROI classifier for semantic class, bounding box, and pixel-wise mask predictions uniquely to Mask R-CNN. Its training optimizes a combined loss comprising object detection and mask prediction losses,

$$\mathcal{L}_{det} = \mathcal{L}_{rpn} + \mathcal{L}_{box} + \mathcal{L}_{mask}, \quad (4.1)$$

where $\mathcal{L}_{rpn}$ penalizes region proposals, $\mathcal{L}_{box}$ guides bounding box regression and classification, and $\mathcal{L}_{mask}$ addresses pixel-wise mask predictions.

This chapter proposes to equip Mask R-CNN with additional domain alignment modules that encourage the model to learn domain-invariant features. These modules operate at three levels: image, instance, and mask.

Image-level Domain Alignment: In FPN-equipped Mask R-CNN, the backbone network generates multi-scale feature maps, and their first level is employed for content

representation purposes during image-level alignment. This representation is subsequently processed through convolutional and fully connected layers for adversarial domain classification.

Let I_n denote the n^{th} of N training images, D_n its corresponding domain label from $|D|$ possible domains, and p_n the $|D|$ -length vector of predicted probabilities of the domain of the image. By employing cross-entropy loss, the loss for image-level domain classification can be expressed as:

$$\mathcal{L}_{dimg} = - \sum_{n=1}^N \mathbf{1}_{D_n} \cdot \log(p_n), \quad (4.2)$$

Here $\mathbf{1}$ denotes the one-hot vector indicator function.

Aligning image-level representations effectively reduces the shift between images by mitigating the influence of global image differences such as style, scale, and illumination [14].

Instance-level Domain Alignment: To address domain shift at the instance level, ROI-based feature vectors processed through fully connected layers are utilized for instance-level domain classification.

The goal of DG is to improve domain-independent parameters across the model. This means that any domain-related information should not be contained in instance-level features. To ensure this, a loss function is included to penalize the presence of domain information in instance features:

$$\mathcal{L}_{dbox} = - \sum_{n=1}^N \sum_{m=1}^M \mathbf{1}_{D_n} \cdot \log(p_{n,m}). \quad (4.3)$$

Here, N and M represent the numbers of images and instances in the training batch, respectively. By aligning the representations at the instance level, we can alleviate differences between local instances, such as variations in object appearance, size, viewpoint, and so on [130].

Pixel-level Domain Alignment: A strategy that focuses on aligning pixel-wise representations across different domains is proposed to address domain shift at the pixel level. The segmentation masks produced by the Mask R-CNN model are leveraged to extract pixel-level features. These features are then processed through additional layers for domain classification at the pixel level.

Similar to instance-level domain alignment, a loss function is included to penalize the presence of domain information in the mask pixels to improve domain-independent parameters at the pixel level across the model:

$$\mathcal{L}_{dmask} = - \sum_{n=1}^N \sum_{m=1}^M \sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{D_n} \cdot \log(p_{n,m,h,w}), \quad (4.4)$$

where N , M , W , and H represent the numbers of images and instances in the training batch and the width and height of the mask features, respectively.

Network Architecture: The proposed network is shown in Fig. 4.2. To train the network through an adversarial approach, a GRL [93] is employed, reversing the loss gradient sign during training. Features are subjected to the GRL before input into the domain classifier network. The total training loss is calculated as the sum of the original Mask R-CNN loss and the newly introduced domain generalization losses:

$$\mathcal{L} = \mathcal{L}_{det} + \alpha \mathcal{L}_{dimg} + \beta \mathcal{L}_{dbox} + \gamma \mathcal{L}_{dmask}, \quad (4.5)$$

where $\mathcal{L}_{det}$ is the Mask R-CNN loss. The hyperparameters α , β , and γ represent the weights assigned to image-level, instance-level, and pixel-level adversarial domain losses and are set empirically to improve performance.

4.3. Experiments

4.3.1. Dataset

In the IEEE DFC 2023 dataset [10], a comprehensive dataset of remote sensing images is provided to identify various building roof types. The images were acquired from the SuperView-1 and Gaofen-2 satellites, each providing spatial resolutions of 0.5 m and 0.8 m, respectively. The dataset comprises images with a spatial size of 512×512 pixels, collected from 17 distinct cities spanning six continents. Each city represents a unique domain. These images have been meticulously annotated, each assigned to one of twelve distinct categories of roof types. However, this dataset was employed for building segmentation for this thesis' research. All 12 distinct categories were treated as a single class, representing buildings, to avoid dealing with the class imbalance issue, as the primary emphasis is on domain generalization.

Table 4.1: Distribution of images and instances across various cities on different continents.

Continent	City	# of images	# of instances
Africa	AddisAbaba	57	4871
	NewDelhi	135	5929
Asia	Suzhou	437	8961
	Tokyo	475	15617
Europe	Barcelona	227	20160
	Berlin	566	22232
	Copenhagen	314	16372
	Portsmouth	155	19362
Austria	Darwin	230	9985
	Sydney	36	951
S. America	Brasilia	248	9864
	Rio	97	8836
	SaoPaulo	116	12384
	SaoLuis	30	9645
N. America	NewYork	198	13528
	Jacksonville	291	6176
	SanDiego	108	9390
Total		3720	194263

4.3.2. Experiment Setup

A three-fold cross-validation experiment was conducted to evaluate the model’s cross-domain performance. In the first fold, three cities from Asia were selected for validation, and four cities from South America were designated for testing. In the second fold, Asian cities were used for validation, with cities from North America used for testing. In the third fold, European cities were used for validation, while Asian cities were used for testing.

Three experiments—a lower baseline, an upper baseline, and domain generalization—

are conducted to compare and evaluate the proposed domain generalization method.

Lower baseline: This experiment uses the standard Mask R-CNN object detection method without applying any domain generalization techniques. The model is optimized based on performance metrics from 80% of the validation set and evaluated on 80% of the test set.

Oracle: This experiment establishes the upper performance bound. Unlike the lower baseline, it incorporates 20% of both the validation and test datasets into the training data.

Domain generalization: This experiment employs the proposed domain generalization method. The neural network is trained on the training dataset, validated on 80% of the validation dataset, and evaluated on 80% of the test dataset. Additionally, results are presented for various configurations involving image-level, instance-level, and pixel-level domain alignments.

4.3.3. Results

The proposed approach was assessed using the mAP metric with both a single threshold (0.5: mAP@50) and multiple thresholds (0.5-0.95: mAP). The experiments were conducted using a three-fold cross-domain setup to ensure the robustness of the domain generalization model.

As presented in Table 4.2, the cross-validation results demonstrate that the proposed method consistently outperforms the baseline Mask R-CNN across different folds. Specifically, when compared against Mask R-CNN (the baseline), the approach achieved average performance gains of +0.96%, +0.63%, and 0.5% using image-level, instance-level, and pixel-level alignments, respectively, with respect to mAP at a 0.5 threshold of Intersection over Union (IoU). However, when combining all three alignment strategies, the performance still surpassed the baseline, albeit slightly lower than the performance achieved with only image-level alignment.

The results indicate that each alignment strategy contributed to the overall performance improvement, demonstrating the effectiveness of domain generalization techniques in remote sensing image analysis. The image-level alignment showed the highest individual improvement, suggesting that variations in the overall image appearance across

Table 4.2: Cross validation experiment results for building detection using different folds with the proposed DGMaskRCNN model. Each fold used images from different continents for training, validation, and testing. The performance metrics reported include mAP and mAP at a 0.5 IoU threshold (mAP@50).

	image	instance	pixel	mAP	mAP@50
Mask R-CNN (lower baseline)					
Fold 1				11.3	23.8
Fold 2				19.4	38.7
Fold 3				15.7	32.3
Average				15.47 $\pm$ 4.06	31.6 $\pm$ 7.47
Mask R-CNN (oracle)					
Fold 1				20.8	42.5
Fold 2				24.3	49
Fold 3				20	39.9
Average				21.7 $\pm$ 2.29	43.8 $\pm$ 4.69
DGMaskRCNN(proposed)					
Fold 1				12	25.4
Fold 2	✓			20.5	40.7
Fold 3				16.8	31.9
Average				16.43 $\pm$ 4.26	32.67 $\pm$ 7.68
DGMaskRCNN(proposed)					
Fold 1				11.9	25.2
Fold 2		✓		19.8	40
Fold 3				16.6	32.1
Average				16.1 $\pm$ 3.97	32.43 $\pm$ 7.41
DGMaskRCNN(proposed)					
Fold 1				11.7	24.2
Fold 2	✓	✓		19.9	40.8
Fold 3				16.3	31.7
Average				15.97 $\pm$ 4.11	32.23 $\pm$ 8.31
DGMaskRCNN(proposed)					
Fold 1				11.4	24.6
Fold 2	✓	✓	✓	20.4	40.7
Fold 3				16.8	31.83
Average				16.2 $\pm$ 4.53	32.38 $\pm$ 8.06

different domains significantly impact the model’s performance. Instance-level alignment also contributed positively, indicating the importance of adjusting for differences at the object level within images from various domains. While contributing the least among the three, Pixel-level alignment still provided a measurable enhancement, highlighting the need to account for fine-grained variations at the pixel level.

When all three alignment strategies were combined, the performance gains were cumulative but not entirely additive, suggesting some level of redundancy or interaction between the methods. Despite this, the combined approach still outperformed the baseline Mask R-CNN, reaffirming the robustness of the proposed domain generalization techniques.

These results underscore the potential of integrating multiple levels of domain alignment to improve model generalization in remote sensing applications. Future work could explore optimizing the balance between these strategies to further enhance performance and investigate other potential alignment techniques or loss functions that might better capture the domain invariances present in remote sensing data.

The findings from this experiment are promising. They show clear improvements in model performance through the application of domain generalization techniques. These advancements contribute to the broader goal of developing more reliable and versatile remote sensing models capable of performing well across diverse and unseen environments.

4.3.4. Ablation Study

The ablation study aims to investigate the impact of varying the loss weights α , β , and γ in Equation 4.5 on the model’s performance, specifically focusing on the mAP and mAP at a 0.5 IoU threshold (mAP@50) metrics.

The results, presented in Table 4.3, demonstrate how different combinations of these weights affect the DGMaskRCNN model’s ability to generalize. The highest performance, with an mAP of 11.5 and mAP@50 of 23.4, was observed when all three weights were set to 0.1. This suggests that a balanced emphasis on the different components of the loss function contributes to optimal performance.

When α is increased to 0.4 while keeping β and γ constant at 0.1, the mAP slightly

drops to 11.4, indicating a marginal decrease in performance. A similar decrease is seen when γ is increased to 0.4 (mAP of 10.3), suggesting that the model’s performance is sensitive to the weighting of these components.

Table 4.3: Results of the ablation study for weights α , β , and γ in Equation 4.5, showing the impact on mAP and mAP at a 0.5 IoU threshold (mAP@50) for the proposed DGMaskRCNN model.

α	β	γ	mAP	mAP@50
0.025	0.1	0.1	10.9	23.9
0.1	0.1	0.1	11.5	23.4
0.4	0.1	0.1	11.4	23
0.1	0.025	0.1	10.5	23
0.1	0.4	0.1	10.4	22.1
0.1	0.1	0.025	11.2	22.9
0.1	0.1	0.4	10.3	22.1

Interestingly, when β is increased to 0.4 with the other weights at 0.1, the model’s mAP decreases more significantly to 10.4, indicating that the balance between these weights is crucial for maintaining performance. Conversely, reducing α or γ to 0.025 results in a noticeable drop in mAP, further emphasizing the importance of maintaining these weights at certain levels for optimal performance.

Overall, the results from this ablation study indicate that the weights in Equation 4.5 play a critical role in determining the model’s effectiveness. Careful tuning of these parameters is essential to achieve the best possible performance for the DGMaskRCNN model.

4.4. Conclusion

This study is the first attempt to equip an instance segmentation model with domain generalization. This has been achieved by adding three domain adversarial modules at image, instance, and mask levels through the use of GRLs. The preliminary results are

promising, as they improve the effectiveness of the standard Mask R-CNN, showing that the method can actually learn domain-invariant features. Future work could explore consistency loss, promoting identical domain predictions at each alignment level and addressing the covariate shift assumption.

5. CONCLUSION

This thesis has introduced novel approaches to address the challenges of domain generalization for instance segmentation and object detection in remote sensing. The research focused on developing and evaluating domain generalization techniques that operate at multiple levels—image, instance, and pixel—integrated with consistency regularization to enhance model performance across diverse domains.

The proposed domain generalization method exhibits substantial advancements over baseline models in object detection. The results indicate that instance-level generalization techniques outperform image-level techniques in buildings-only datasets, while image-level generalizations are more effective in scenarios with varied roof types. The combination of instance-level and image-level techniques, further reinforced by consistency regularization, leads to the highest performance gains. Specifically, in the buildings-only setting, the integration of these techniques achieves the best mAP and mAP@50 scores, demonstrating the robustness of the approach.

For instance segmentation, the study introduces a domain generalized Mask R-CNN model equipped with domain adversarial modules at the image, instance, and mask levels. The results from the IEEE Data Fusion Contest 2023 dataset demonstrate improvements in model performance through the application of these domain generalization techniques. The image-level alignment strategy contributes the most to performance enhancement, followed by instance-level and pixel-level alignments. The combined approach of these strategies, although not entirely additive, still outperforms the baseline, reaffirming the efficacy of the proposed techniques.

The results obtained from experiments in object detection and instance segmentation highlight the effectiveness of multi-level domain alignment strategies in enhancing the generalization ability of models for remote sensing tasks. These advancements contribute to the broader goal of developing more reliable and versatile remote sensing models capable of performing well across diverse and unseen environments.

Future work can explore optimizing the balance between alignment strategies and investigating additional techniques or loss functions that better capture domain invariances. Moreover, applying consistency loss to promote identical domain predictions at each

alignment level and addressing the covariate shift assumption could further enhance the robustness of these models.

In conclusion, this thesis has made significant contributions to the field of domain generalization in remote sensing, providing a foundation for future research to build upon and further improve the generalization capabilities of remote sensing models.

REFERENCES

- [1] Singh B., Jeganathan C., Rathore V., (n.d.), “Remote sensing technology for monitoring and modelling ecological processes”.
- [2] Girshick R., Donahue J., Darrell T., Malik J., (2014), Rich feature hierarchies for accurate object detection and semantic segmentation, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, 580–587.
- [3] Girshick R., (2015), Fast r-cnn, in “Proceedings of the IEEE international conference on computer vision”, 1440–1448.
- [4] Ren S., He K., Girshick R., Sun J., (2015), “Faster r-cnn: Towards real-time object detection with region proposal networks”, *Advances in neural information processing systems*.
- [5] Redmon J., Divvala S., Girshick R., Farhadi A., (2016), You only look once: Unified, real-time object detection, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, 779–788.
- [6] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A. C., (2016), Ssd: Single shot multibox detector, in “Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14”, Springer, 21–37.
- [7] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S., (2020), End-to-end object detection with transformers, in “European conference on computer vision”, Springer, 213–229.
- [8] He K., Gkioxari G., Dollár P., Girshick R., (2017), Mask r-cnn, in “Proceedings of the IEEE international conference on computer vision”, 2961–2969.
- [9] Li D., Yang Y., Song Y.-Z., Hospedales T. M., (2017), Deeper, broader and artier domain generalization, in “Proceedings of the IEEE international conference on computer vision”, 5542–5550.
- [10] Persello C., Hänsch R., Vivone G., Chen K., Yan Z., Tang D., Huang H., Schmitt M., Sun X., (2023), “2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction [technical committees]”, *IEEE Geoscience and Remote Sensing Magazine*, 11(1), 94–97.
- [11] Aptoula E., Ariman S., (2020), “Hierarchical spatial–spectral features for the chlorophyll-a estimation of lake balik, turkey”, *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.
- [12] Mao Y., Chen K., Zhao L., Chen W., Tang D., Liu W., Wang Z., Diao W., Sun X., Fu K., (2023), “Elevation estimation-driven building 3d reconstruction from single-view remote sensing imagery”, *IEEE Transactions on Geoscience and Re-*

remote Sensing.

- [13] Sarkar A., Chowdhury T., Murphy R. R., Gangopadhyay A., Rahnemounfar M., (2023), “Sam-vqa: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery”, *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–16.
- [14] Seemakurthy K., Fox C., Aptoula E., Bosilj P., (2023), Domain generalised faster r-cnn, in “Proceedings of the AAAI Conference on Artificial Intelligence”, 2180–2190.
- [15] Liu Q., Ren K., Meng X., Shao F., (2023), “Domain adaptive cross-reconstruction for change detection of heterogeneous remote sensing images via a feedback guidance mechanism”, *IEEE Transactions on Geoscience and Remote Sensing*.
- [16] Gong H., Mu T., Li Q., Dai H., Li C., He Z., Wang W., Han F., Tuniyazi A., Li H., et al. (2022), “Swin-transformer-enabled yolov5 with attention mechanism for small object detection on satellite images”, *Remote Sensing*, 14(12), 2861.
- [17] Xia G.-S., Bai X., Ding J., Zhu Z., Belongie S., Luo J., Datcu M., Pelillo M., Zhang L., (2018), Dota: A large-scale dataset for object detection in aerial images, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, 3974–3983.
- [18] Wang K., Fu X., Huang Y., Cao C., Shi G., Zha Z.-J., (2023), Generalized uav object detection via frequency domain disentanglement, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, 1064–1073.
- [19] Shamsolmoali P., Chanussot J., Zareapoor M., Zhou H., Yang J., (2021), “Multi-patch feature pyramid network for weakly supervised object detection in optical remote sensing images”, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13.
- [20] Gong Y., Zhang Z., Wen J., Lan G., Xiao S., (2023), “Small ship detection of sar images based on optimized feature pyramid and sample augmentation”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- [21] Zhang L., Zhang L., Tao D., Huang X., (2013), “Sparse transfer manifold embedding for hyperspectral target detection”, *IEEE Transactions on Geoscience and Remote Sensing*, 52(2), 1030–1043.
- [22] Luo H., Feng X., Du B., Zhang Y., (2024), “A multimodal feature fusion network for building extraction with very high-resolution remote sensing image and lidar data”, *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–19.
- [23] Li P., Chen X., Shen S., (2019), Stereo r-cnn based 3d object detection for autonomous driving, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, 7644–7652.

- [24] Zhao Y., Sun G., Zhang L., Zhang A., Jia X., Han Z., (2023), “Msrf-net: multi-scale receptive field network for building detection from remote sensing images”, *IEEE Transactions on Geoscience and Remote Sensing*.
- [25] Xiang B., Peters T., Kontogianni T., Vetterli F., Puliti S., Astrup R., Schindler K., (2023), “Towards accurate instance segmentation in large-scale lidar point clouds”, *arXiv preprint arXiv:2307.02877*.
- [26] Yasir M., Zhan L., Liu S., Wan J., Hossain M. S., Isiacik Colak A. T., Liu M., Islam Q. U., Raza Mehdi S., Yang Q., (2023), “Instance segmentation ship detection based on improved yolov7 using complex background sar images”, *Frontiers in Marine Science*, 10, 1113669.
- [27] Zurowietz M., Nattkemper T. W., (2020), “Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration”, *IEEE Access*, 8, 143558–143568.
- [28] Liang C., Cheng B., Xiao B., Dong Y., Chen J., (2023), “Multilevel heterogeneous domain adaptation method for remote sensing image segmentation”, *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–16.
- [29] Zhao Z.-Q., Zheng P., Xu S.-t., Wu X., (2019), “Object detection with deep learning: A review”, *IEEE transactions on neural networks and learning systems*, 30(11), 3212–3232.
- [30] Uijlings J. R., Van De Sande K. E., Gevers T., Smeulders A. W., (2013), “Selective search for object recognition”, *International journal of computer vision*, 104, 154–171.
- [31] Lowe D. G., (2004), “Distinctive image features from scale-invariant keypoints”, *International journal of computer vision*, 60, 91–110.
- [32] Dalal N., Triggs B., (2005), Histograms of oriented gradients for human detection, in “2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)”, Vol. 1, Ieee, 886–893.
- [33] Lienhart R., Maydt J., (2002), An extended set of haar-like features for rapid object detection, in “Proceedings. international conference on image processing”, Vol. 1, IEEE, I–I.
- [34] Cortes C., Vapnik V., (1995), “Support vector machine”, *Machine learning*, 20(3), 273–297.
- [35] Freund Y., Schapire R. E., (1997), “A decision-theoretic generalization of on-line learning and an application to boosting”, *Journal of computer and system sciences*, 55(1), 119–139.
- [36] Felzenszwalb P. F., Girshick R. B., McAllester D., Ramanan D., (2009), “Object detection with discriminatively trained part-based models”, *IEEE transactions on*

pattern analysis and machine intelligence, 32(9), 1627–1645.

- [37] Lin T.-Y., Goyal P., Girshick R., He K., Dollár P., (2017), Focal loss for dense object detection, in “Proceedings of the IEEE international conference on computer vision”, 2980–2988.
- [38] Tan M., Pang R., Le Q. V., (2020), Efficientdet: Scalable and efficient object detection, in “Proceedings of the IEEE/CVF conference on computer vision and pattern recognition”, 10781–10790.
- [39] Gui S., Song S., Qin R., Tang Y., (2024), “Remote sensing object detection in the deep learning era—a review”, *Remote Sensing*, 16(2), 327.
- [40] Jaderberg M., Simonyan K., Zisserman A., et al. (2015), “Spatial transformer networks”, *Advances in neural information processing systems*.
- [41] Chen L.-C., Hermans A., Papandreou G., Schroff F., Wang P., Adam H., (2018), Masklab: Instance segmentation by refining object detection with semantic and direction features, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, 4013–4022.
- [42] Lin T., Dollár P., Girshick R., He K., Hariharan B., Belongie S., (2017), Feature pyramid networks for object detection, in “IEEE CVPR”, 2117–2125.
- [43] Cai Z., Vasconcelos N., (2019), “Cascade r-cnn: High quality object detection and instance segmentation”, *IEEE transactions on pattern analysis and machine intelligence*, 1483–1498.
- [44] Chen K., Pang J., Wang J., Xiong Y., Li X., Sun S., Feng W., Liu Z., Shi J., Ouyang W., et al. (2019), Hybrid task cascade for instance segmentation, in “Proceedings of the IEEE/CVF conference on computer vision and pattern recognition”, 4974–4983.
- [45] Arnab A., Torr P. H., (2017), Pixelwise instance segmentation with a dynamically instantiated network, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, 441–450.
- [46] Li Q., Arnab A., Torr P. H., (2018), Weakly-and semi-supervised panoptic segmentation, in “Proceedings of the European conference on computer vision (ECCV)”, 102–118.
- [47] Gao N., Shan Y., Wang Y., Zhao X., Yu Y., Yang M., Huang K., (2019), Ssap: Single-shot instance segmentation with affinity pyramid, in “Proceedings of the IEEE/CVF international conference on computer vision”, 642–651.
- [48] Sun X., Wang H., Fu K., (2009), “Automatic detection of geospatial objects using taxonomic semantics”, *IEEE Geoscience and Remote Sensing Letters*, 7(1), 23–27.
- [49] Lin Y., He H., Yin Z., Chen F., (2014), “Rotation-invariant object detection in

remote sensing images based on radial-gradient angle”, IEEE Geoscience and Remote Sensing Letters, 12(4), 746–750.

- [50] Leninisha S., Vani K., (2015), “Water flow based geometric active deformable model for road network”, ISPRS Journal of Photogrammetry and Remote Sensing, 102, 140–147.
- [51] Chaudhuri D., Samal A., (2008), “An automatic bridge detection technique for multispectral images”, IEEE transactions on geoscience and remote sensing, 46(9), 2720–2727.
- [52] Cheng G., Han J., Zhou P., Guo L., (2014), “Multi-class geospatial object detection and geographic image classification based on collection of part detectors”, ISPRS Journal of Photogrammetry and Remote Sensing, 98, 119–132.
- [53] Yao Y., Cheng G., Wang G., Li S., Zhou P., Xie X., Han J., (2022), “On improving bounding box representations for oriented object detection”, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–11.
- [54] Huang Z., Li W., Xia X.-G., Wu X., Cai Z., Tao R., (2021), “A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images”, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–20.
- [55] Shamsolmoali P., Zareapoor M., Chanussot J., Zhou H., Yang J., (2021), “Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery”, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14.
- [56] Zhang X., Zhang T., Wang G., Zhu P., Tang X., Jia X., Jiao L., (2023), “Remote sensing object detection meets deep learning: A metareview of challenges and advances”, IEEE Geoscience and Remote Sensing Magazine.
- [57] Zhou K., Liu Z., Qiao Y., Xiang T., Loy C. C., (2022), “Domain generalization: A survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [58] Wang J., Lan C., Liu C., Ouyang Y., Qin T., Lu W., Chen Y., Zeng W., Yu P., (2022), “Generalizing to unseen domains: A survey on domain generalization”, IEEE Transactions on Knowledge and Data Engineering.
- [59] Caruana R., (1997), “Multitask learning”, Machine learning, 28, 41–75.
- [60] Mallya A., Davis D., Lazebnik S., (2018), Piggyback: Adapting a single network to multiple tasks by learning to mask weights, in “Proceedings of the European conference on computer vision (ECCV)”, 67–82.
- [61] Liu S., Johns E., Davison A. J., (2019), End-to-end multi-task learning with attention, in “Proceedings of the IEEE/CVF conference on computer vision and pattern recognition”, 1871–1880.

- [62] Wang S., Yu L., Li C., Fu C.-W., Heng P.-A., (2020), Learning from extrinsic and intrinsic supervisions for domain generalization, in “European Conference on Computer Vision”, Springer, 159–176.
- [63] Albuquerque I., Naik N., Li J., Keskar N., Socher R., (2020), “Improving out-of-distribution generalization via multi-task self-supervised pretraining”, arXiv preprint arXiv:2003.13525.
- [64] Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., Xiong H., He Q., (2020), “A comprehensive survey on transfer learning”, *Proceedings of the IEEE*, 109(1), 43–76.
- [65] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., (2009), Imagenet: A large-scale hierarchical image database, in “2009 IEEE conference on computer vision and pattern recognition”, Ieee, 248–255.
- [66] Wang M., Deng W., (2018), “Deep visual domain adaptation: A survey”, *Neuro-computing*, 312, 135–153.
- [67] Oza P., Sindagi V. A., Sharmini V. V., Patel V. M., (2023), “Unsupervised domain adaptation of object detectors: A survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [68] Ismael S. F., Kayabol K., Aptoula E., (2023), “Unsupervised domain adaptation for the semantic segmentation of remote sensing images via one-shot image-to-image translation”, *IEEE Geoscience and Remote Sensing Letters*.
- [69] Chen J., Zhu J., Guo Y., Sun G., Zhang Y., Deng M., (2022), “Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention”, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.
- [70] Hoffman J., Tzeng E., Park T., Zhu J.-Y., Isola P., Saenko K., Efros A., Darrell T., (2018), Cycada: Cycle-consistent adversarial domain adaptation, in “International conference on machine learning”, Pmlr, 1989–1998.
- [71] Honarvar Nazari N., Kovashka A., (2020), Domain generalization using shape representation, in “Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16”, Springer, 666–670.
- [72] Yue X., Zhang Y., Zhao S., Sangiovanni-Vincentelli A., Keutzer K., Gong B., (2019), Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data, in “Proceedings of the IEEE/CVF International Conference on Computer Vision”, 2100–2110.
- [73] Prakash A., Boochoon S., Brophy M., Acuna D., Cameracci E., State G., Shapira O., Birchfield S., (2019), Structured domain randomization: Bridging the reality gap by context-aware synthetic data, in “2019 International Conference on Robotics and Automation (ICRA)”, IEEE, 7249–7255.

- [74] Huang J., Guan D., Xiao A., Lu S., (2021), Fsdrr: Frequency space domain randomization for domain generalization, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, 6891–6902.
- [75] Khiredkar R., Yoo D., Kitani K., (2019), Domain randomization for scene-specific car detection and pose estimation, in “2019 IEEE Winter Conference on Applications of Computer Vision (WACV)”, IEEE, 1932–1940.
- [76] Shankar S., Piratla V., Chakrabarti S., Chaudhuri S., Jyothi P., Sarawagi S., (2018), “Generalizing across domains via cross-gradient training”, arXiv preprint arXiv:1804.10745.
- [77] Volpi R., Namkoong H., Sener O., Duchi J. C., Murino V., Savarese S., (2018), “Generalizing to unseen domains via adversarial data augmentation”, Advances in neural information processing systems.
- [78] Zhou K., Yang Y., Hospedales T., Xiang T., (2020), Deep domain-adversarial image generation for domain generalisation, in “Proceedings of the AAAI conference on artificial intelligence”, 13025–13032.
- [79] Blanchard G., Deshmukh A. A., Dogan Ü., Lee G., Scott C., (2021), “Domain generalization by marginal transfer learning”, The Journal of Machine Learning Research, 22(1), 46–100.
- [80] Li Y., Gong M., Tian X., Liu T., Tao D., (2018), Domain generalization via conditional invariant representations, in “Proceedings of the AAAI conference on artificial intelligence”.
- [81] Hu S., Zhang K., Chen Z., Chan L., (2020), Domain generalization via multidomain discriminant analysis, in “Uncertainty in Artificial Intelligence”, PMLR, 292–302.
- [82] Erfani S., Baktashmotlagh M., Moshtaghi M., Nguyen X., Leckie C., Bailey J., Kotagiri R., (2016), Robust domain generalisation by enforcing distribution invariance, in “Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)”, AAAI Press, 1455–1461.
- [83] Jin X., Lan C., Zeng W., Chen Z., (2020), “Feature alignment and restoration for domain generalization and adaptation”, arXiv preprint arXiv:2006.12009.
- [84] Motiian S., Piccirilli M., Adjeroh D. A., Doretto G., (2017), Unified deep supervised domain adaptation and generalization, in “Proceedings of the IEEE international conference on computer vision”, 5715–5725.
- [85] Mahajan D., Tople S., Sharma A., (2021), Domain generalization using causal matching, in “International Conference on Machine Learning”, PMLR, 7313–7324.
- [86] Zhao S., Gong M., Liu T., Fu H., Tao D., (2020), “Domain generalization via

entropy regularization”, *Advances in Neural Information Processing Systems*, 33, 16096–16107.

- [87] Wang Z., Loog M., van Gemert J., (n.d.), “Respecting domain relations: Hypothesis invariance for domain generalization. in 2020 25th international conference on pattern recognition (icpr) 2021 jan 10 (pp. 9756-9763)”.
- [88] Tzeng E., Hoffman J., Zhang N., Saenko K., Darrell T., (2014), “Deep domain confusion: Maximizing for domain invariance”, *arXiv preprint arXiv:1412.3474*.
- [89] Wang J., Chen Y., Feng W., Yu H., Huang M., Yang Q., (2020), “Transfer learning with dynamic distribution adaptation”, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1), 1–25.
- [90] Matsuura T., Harada T., (2020), Domain generalization using a mixture of multiple latent domains, in “Proceedings of the AAAI Conference on Artificial Intelligence”, 11749–11756.
- [91] Akuzawa K., Iwasawa Y., Matsuo Y., (2020), Adversarial invariant feature learning with accuracy constraint for domain generalization, in “Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II”, Springer, 315–331.
- [92] Li Y., Tian X., Gong M., Liu Y., Liu T., Zhang K., Tao D., (2018), Deep domain generalization via conditional invariant adversarial networks, in “Proceedings of the European conference on computer vision (ECCV)”, 624–639.
- [93] Ganin Y., Lempitsky V., (2015), Unsupervised domain adaptation by backpropagation, in “International conference on machine learning”, PMLR, 1180–1189.
- [94] Rahman M. M., Fookes C., Baktashmotlagh M., Sridharan S., (2020), “Correlation-aware adversarial domain adaptation and generalization”, *Pattern Recognition*, 100, 107124.
- [95] Ilse M., Tomczak J. M., Louizos C., Welling M., (2020), Diva: Domain invariant variational autoencoders, in “Medical Imaging with Deep Learning”, PMLR, 322–348.
- [96] Wang G., Han H., Shan S., Chen X., (2020), Cross-domain face presentation attack detection via multi-domain disentangled representation learning, in “Proceedings of the IEEE/CVF conference on computer vision and pattern recognition”, 6678–6687.
- [97] Chattopadhyay P., Balaji Y., Hoffman J., (2020), Learning to balance specificity and invariance for in and out of domain generalization, in “Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16”, Springer, 301–318.

- [98] Piratla V., Netrapalli P., Sarawagi S., (2020), Efficient domain generalization via common-specific low-rank decomposition, in “International Conference on Machine Learning”, PMLR, 7728–7738.
- [99] Qi L., Wang L., Shi Y., Geng X., (2021), “Unsupervised domain generalization for person re-identification: A domain-specific adaptive framework”, arXiv preprint arXiv:2111.15077.
- [100] Zhou K., Yang Y., Qiao Y., Xiang T., (2021), “Domain adaptive ensemble learning”, IEEE Transactions on Image Processing, 30, 8008–8018.
- [101] Ding Z., Fu Y., (2017), “Deep domain generalization with structured low-rank constraint”, IEEE Transactions on Image Processing, 27(1), 304–313.
- [102] Cha J., Cho H., Lee K., Park S., Lee Y., Park S., (2021), “Domain generalization needs stochastic weight averaging for robustness on domain shifts”, arXiv preprint arXiv:2102.08604, 3, 3.
- [103] Santoro A., Bartunov S., Botvinick M., Wierstra D., Lillicrap T., (2016), Meta-learning with memory-augmented neural networks, in “International conference on machine learning”, PMLR, 1842–1850.
- [104] Finn C., Abbeel P., Levine S., (2017), Model-agnostic meta-learning for fast adaptation of deep networks, in “International conference on machine learning”, PMLR, 1126–1135.
- [105] Snell J., Swersky K., Zemel R., (2017), “Prototypical networks for few-shot learning”, Advances in neural information processing systems.
- [106] Zhu S., Wu C., Du B., Zhang L., (2023), “Style and content separation network for remote sensing image cross-scene generalization”, ISPRS Journal of Photogrammetry and Remote Sensing, 201, 1–11.
- [107] Luo M., Ji S., Wei S., (2023), “A diverse large-scale building dataset and a novel plug-and-play domain generalization method for building extraction”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
- [108] Zhang Y., Zhang M., Li W., Wang S., Tao R., (2023), “Language-aware domain generalization network for cross-scene hyperspectral image classification”, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–12.
- [109] Gao Z., Pan B., Xu X., Li T., Shi Z., (2023), “Lica: Label-indicate-conditional-alignment domain generalization for pixel-wise hyperspectral imagery classification”, IEEE Transactions on Geoscience and Remote Sensing.
- [110] Zheng J., Wu W., Yuan S., Fu H., Li W., Yu L., (2021), “Multisource-domain generalization-based oil palm tree detection using very-high-resolution (vhr) satellite images”, IEEE Geoscience and Remote Sensing Letters, 19, 1–5.
- [111] Zhao Q., Wang X., Wang B., Wang L., Liu W., Li S., (2023), “A dual-attention

deep discriminative domain generalization model for hyperspectral image classification”, *Remote Sensing*, 15(23), 5492.

- [112] Liang C., Li W., Dong Y., Fu W., (2024), “Single domain generalization method for remote sensing image segmentation via category consistency on domain randomization”, *IEEE Transactions on Geoscience and Remote Sensing*.
- [113] Wang B., Xu Y., Wu Z., Zheng S., Wei Z., Chanussot J., (2024), “Hyperspectral images single-source domain generalization based on nonlinear sample generation”, *IEEE Transactions on Geoscience and Remote Sensing*.
- [114] Iizuka R., Xia J., Yokoya N., (2023), “Frequency-based optimal style mix for domain generalization in semantic segmentation of remote sensing images”, *IEEE Transactions on Geoscience and Remote Sensing*.
- [115] Qi Y., Zhang J., Liu D., Zhang Y., (2024), “Multi-source domain generalization two-branch network for hyperspectral image cross-domain classification”, *IEEE Geoscience and Remote Sensing Letters*.
- [116] Cheng G., Han J., (2016), “A survey on object detection in optical remote sensing images”, *ISPRS journal of photogrammetry and remote sensing*, 117, 11–28.
- [117] Li K., Wan G., Cheng G., Meng L., Han J., (2020), “Object detection in optical remote sensing images: A survey and a new benchmark”, *ISPRS journal of photogrammetry and remote sensing*, 159, 296–307.
- [118] Sun C. P. R. H. G. V. K. C. Z. Y. D. T. H. H. M. S. X., (2022), “2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction”.
- [119] Zhang S., Cao Y., Sui B., (2023), “DF-Mask R-CNN: Direction field-based optimized instance segmentation network for building instance extraction”, *IEEE GRSL*.
- [120] de Carvalho O., de Carvalho J., et al. (2022), “Bounding box-free instance segmentation using semi-supervised iterative learning for vehicle detection”, *IEEE JSTARS*, 15, 3403–3420.
- [121] Wei S., Zeng X., et al. (2022), “LFG-Net: Low-level feature guided network for precise ship instance segmentation in sar images”, *IEEE TGRS*.
- [122] Shi F., Zhang T., (2022), “An anchor-free network with box refinement and saliency supplement for instance segmentation in remote sensing images”, *IEEE GRSL*.
- [123] Kirillov A., Levinkov E., Andres B., Savchynskyy B., Rother C., (2017), Instancecut: from edges to instances with multicut, in “*IEEE CVPR*”, 5008–5017.
- [124] Xie E., Sun P., et al. (2020), Polarmask: Single shot instance segmentation with polar representation, in “*ECCV*”, 12193–12202.

- [125] Fang L., Jiang Y., Yan Y., Yue J., Deng Y., (2023), “Hyperspectral image instance segmentation using spectral–spatial feature pyramid network”, IEEE TGRS.
- [126] Liu G., Peng B., et al. (2023), Fine-grained building roof instance segmentation based on domain adapted pretraining and composite dual-backbone, in “IEEE IGARSS”, 670–673.
- [127] Kang J., Wang Z., et al. (2022), “Disoptnet: Distilling semantic knowledge from optical images for weather-independent building segmentation”, IEEE TGRS.
- [128] Duraklı E., Aptoula E., (2023), Domain generalized object detection for remote sensing images, in “IEEE SIU”.
- [129] He K., Zhang X., Ren S., Sun J., (2016), Deep residual learning for image recognition, in “IEEE CVPR”, 770–778.
- [130] Seemakurthy K., Bosilj P., Aptoula E., Fox C., (2023), Domain generalised fully convolutional one stage detection, in “IEEE ICRA”.

BIOGRAPHY

Efkan DURAKLI received his Bachelor's degree from the Computer Engineering Department of Gebze Technical University in June 2019. He is currently pursuing a Master's degree in Computer Engineering at Gebze Technical University and works as a research assistant at the Institute of Information Technologies at the same university.

PUBLICATIONS WITHIN the SCOPE of THESIS STUDY

E. Durakli, E. Aptoula, Domain generalized object detection for remote sensing images, 31. IEEE Signal Processing and Communication Applications Conference, Istanbul, Türkiye, July, 2023

E. Durakli, P. Bosilj, C. Fox, E. Aptoula, A domain generalized Mask R-CNN for building instance segmentation, IEEE Geoscience and Remote Sensing Symposium (IGARSS), Athens, Greece, July, 2024