Deep Learning Neural Networks to Improve Small Object Detection

UNIVERSIDAD DE MÁLAGA

PhD. THESIS

Iván García Aguilar

Tesis Doctoral por Compendio de Publicaciones
Programa de Doctorado en Tecnologías Informáticas
Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

Noviembre 2023

Documento maquetado con TEXIS v.1.0+.

Deep Learning Neural Networks to Improve Small Object Detection

Memorandum for obtaining the PhD. degree by the University of Málaga presented by

Iván García Aguilar

Directed by

Rafael Marcos Luque Baena PhD. and Ezequiel López Rubio PhD.

Tesis Doctoral por Compendio de Publicaciones
Programa de Doctorado en Tecnologías Informáticas
Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

Noviembre 2023

AUTOR: Iván García Aguilar
https://orcid.org/0000-0001-5476-6704

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña IVÁN GARCÍA AGUILAR
Estudiante del programa de doctorado TECNOLOGÍAS INFORMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: Deep Learning Neural Networks to Improve Small Object Detection.

Realizada bajo la tutorización de RAFAEL MARCOS LUQUE BAENA y dirección de RAFAEL MARCOS LUQUE BAENA Y EZEQUIEL LÓPEZ RUBIO (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 22 de Noviembre de 2023

Fdo.: IVÁN GARCÍA AGUILAR Doctorando/a	Fdo.: RAFAEL MARCOS LUQUE BAENA Tutor/a

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Fdo.: RAFAEL MARCOS LUQUE BAENA Y EZEQUIEL LÓPEZ RUBIO Director/es de tesis

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

UNIVERSIDAD DE MÁLAGA

ANDALUCÍA TECH
Campus de Excelencia Internacional

AUTORIZACIÓN PARA LA LECTURA E INFORME SOBRE LA TESIS DE D. IVÁN GARCÍA AGUILAR

Rafael Marcos Luque Baena, Profesor Titular del departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, en calidad de tutor y director de la tesis doctoral de D. Iván García Aguilar titulada **Deep Learning Neural Networks to Improve Small Object Detection**; y Ezequiel López Rubio, Catedrático de Universidad del departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, en calidad de director de dicha tesis, AUTORIZAN su lectura.

Asimismo, Rafael Marcos Luque Baena y Ezequiel López Rubio, en calidad de tutor y directores de la mencionada tesis, INFORMAN que las publicaciones que avalan la tesis no han sido utilizadas en tesis anteriores. También señalan la idoneidad de presentar la tesis por compendio de artículos dada la gran cantidad de producción científica de alta calidad en problemas iguales o relacionados.

Málaga a 22 de Noviembre de 2023.

Fdo: Rafael Marcos Luque Baena

Fdo: Ezequiel López Rubio

To my family for all the support I have received and to my anxiety for challenging and strengthening me.

The future belongs to those who believe in the beauty of their dreams.

Eleanor Roosevelt.

Acknowledgments

The only way to do great work is to love what you do.
Steve Jobs.

At the beginning of this acknowledgments section, as it should be, I would like to thank my family, especially my parents, Francisco and Paqui. They have been my supporters since childhood and have been by my side in all my decisions. They have taught me all the values necessary to face any challenge. I would also like to mention my brother, Carlos. I do not forget the rest of my family, especially my grandparents, who have always been happy with all my achievements.

Secondly, I would like to thank all those involved in the ICAI (Computational Intelligence and Image Analysis) Working Group. I want to thank my thesis advisors. To Rafael, I am grateful for the proposal offered at the beginning of this Ph.D. program. To Ezequiel, I thank him for the inspiration and knowledge that made this thesis possible. With the help of both of them and their hints, it was possible to complete this thesis agile and efficient. I can't forget my teammates Enrique, Esteban, Miguel Ángel, Juan Miguel, Karl, Jesús, Rosa, Ariadna, José David, and Jorge. Special mention to the latter for being my guide in the first steps of this journey and for all his help.

It is hard to believe that what started as a faraway dream has become a reality today. It has been a journey that has had its ups and downs, but without a doubt, I would do it all over again.

Contents

List of Figures

Acronyms

AAW	Adaptive Anomaly Weight
AI	Artificial Intelligence
BBOX	Bounding Box
BCE	Binary Cross Entropy
CARLA	Car Learning to Act
CBA	Compact Bat Algorithm
CNN	Convolutional Neural Networks
CV	Computer Vision
DA	Data Augmentation
DCNN	Deep Convolutional Neural Networks
DEEPSORT	Deep Simple Online Realtime Tracking
DenseNet	Densely Connected Networks
DL	Deep Learning
DLADT-PW	Deep Learning based Anomaly Detection Technique in Pedestrian Walkways
DNN	Deep Neural Networks
DoPB	Dynamics of Pedestrian Behavior
EDSR	Enhanced Deep Super-resolution Network
ESPCN	Efficient Sub-Pixel Convolutional Neural Network
ESRGAN	Enhanced Super-Resolution Generative Adversarial Networks
EX	Expert Systems
FASTERCNN	Faster Region-Convolutional Neural Network
FASTRCNN	Fast Region-based Convolutional Network
FPN	Feature Pyramid Network
FSRCNN	Fast Super-Resolution Convolutional Neural Network
GAN	Generative Adversarial Networks

GT ..... Ground Truth
HOG ..... Histogram of Oriented Gradients
ICAE ..... Integrated Computer-Aided Engineering
ICAI ..... Computational Intelligence and Image Analysis
ILSVRC ..... ImageNet Large Scale Visual Recognition Challenge
IWANN ..... International Work Conference on Artificial Neural Networks
IWINAC ..... International Work-Conference on the Interplay Between Natural and Artificial Computation
JCR ..... Journal Citation Report
LAPSRN ..... Laplacian Pyramid Super-Resolution Network
LSH ..... Locality Sensitive Hashing
LSOF ..... Large Scale Optical Flow
LSTM ..... Long short-term memory
LUT ..... Look-Up Tables
MAP ..... Mean Average Precision
MASKRCNN . Mask Region-based Convolutional Neural Network
MISVM ..... Multiple Instance Support Vector Machine
MP ..... Max Pooling
MRRCNN .... Multi-Region CNN
MTGAN ..... Generative Adversarial Network
NGSIM ..... Next Generation Simulation
NMS ..... Non Maximum Suppression
PFPN ..... Panoptic Feature Pyramid Network
PFPN-ADT .. Panoptic Feature Pyramid Network-based Anomaly Detection and Tracking
PRL ..... Pattern Recognition Letters
PTZ ..... Pan-Tilt-Zoom
RCNN ..... Regions with CNN Features
REALESRGAN Real Enhanced Super Resolution Generative Adversarial Network
RELU ..... Rectified Linear Unit Function
RGB ..... Red-Green-Blue
RPN ..... Region Proposal Network
SAE ..... Stacked Auto Encoder
SAHI ..... Slicing Aided Hyper Inference

SGD ..... Stochastic Gradient Descent
SIFT ..... Scale Invariant Feature Transform
SOD ..... Small Object Detection
SR ..... Super-resolution
SRCNN ..... Super-Resolution Convolutional Neural Network
SRGAN ..... Super-Resolution Generative Adversarial Networks
SURF ..... Speeded Up Robust Features
YOLACT++ . You Only Look At CoefficientTs ++
YOLO ..... You Only Look Once

Abstract

Success is walking from failure to failure
with no loss of enthusiasm.

Winston Churchill.

Nowadays, the dizzying proliferation of systems for generating multimedia content has led to an **exponential increase** in the **data collected**, allowing several advances in AI (Artificial Intelligence) and CV (Computer Vision) in the last decade. Given this increase in the amount of information, it is a **challenge** to analyze it manually because it is an impossible task due to the time required or by classical methods due to their shortcomings. New techniques are needed to accomplish this task. The field of CV based on CNN (Convolutional Neural Networks), particularly DL (Deep Learning), is the solution to overcome the current limitations in this field.

The main problem addressed in this Ph.D. thesis is the **improvement of the detection of small objects** in road video sequences using CV and SR (Super-resolution). The goal is to significantly enhance the input image to improve the inference and processing using an object detection or segmentation model. As a derived problem, anomaly detection is addressed using these tools. The approach based on improving the detection of small objects in road video sequences is a critical area of research for several reasons. Current advances in object detection, such as ImageNet-trained models, have demonstrated good performance in identifying large and prominent objects. However, they still face several challenges in the area of small objects, which are common in traffic scenarios. Therefore, this line of research represents a relevant and challenging problem within the field.

As a witness of these objectives, a series of research works performed during the thesis are included. The first one was published in the journal Expert Systems in 2021 and discussed the improvement of the detection of small objects using SR and re-inference techniques to improve the accuracy obtained by object detection models. The proposal consistently outperforms pre-trained models, achieving an average accuracy of 45.1% with the EfficientDet-D4 model for the initial video sequence, a significant improvement over the original model that obtained a 24.3%. The second paper was presented at the IWINAC (International Work-Conference on the Interplay Between Natural and Artificial Computation) in 2022 and proposed applying these techniques in object segmentation environments applied to urban sequences. The presented approach demonstrates a notable improvement in several objects. For example, the Person class increased from 89.1% to 93.5%, and the item detection rate improved from 42.71% to 71.53%. The third paper was

published in the journal ICAE (Integrated Computer-Aided Engineering) in 2023 and deals with applying the previously described techniques in object segmentation in urban sequences, using heuristics to optimize the performance of the applied methodology. Experimental results validate the proposal's effectiveness, showing up to an 8.1% improvement with the YOLACT++ model in a sequence benefiting instant segmentation in surveillance and transportation systems. The fourth paper was published in PRL (Pattern Recognition Letters) in 2023 and is based on developing a methodology that allows automatic training by fine-tuning an object recognition model without manual annotation to improve the identification of small elements. Metrics obtained by our proposal surpass the original model. For example, in sequence two, the average mAP increases from 27.7% to 92.6%, consistently improving accuracy in several domains. The fifth paper was presented at the IWANN (International Work Conference on Artificial Neural Networks) in 2023 and is based on the minimum generation of zones of interest in a sequence according to the density calculation of the required elements based on a computed matrix. The presented methodology significantly improves mAP by up to 22.7% compared to the RAW model across test sequences, detecting more objects without requiring retraining. The sixth paper was presented in the journal Neural Computing and Applications in 2023 and deals with the implementation of optimization techniques using a graph algorithm together with SR techniques to relatively reduce the number of sub-images on which to re-infer, thus significantly improving the processing time. Compelling results show a 44.6% increase in detection rates, transitioning from 14.5% to 59.1% with the EfficientDet D4 model using the proposed methodology in the first sequence. Finally, the latest paper of this thesis was published in Sensors 2023, where a methodology for detecting anomalies in synthetic urban sequences is presented. This methodology includes image enhancement techniques to improve the detections of the selected neural network. For example, using the presented methodology, EfficientDet D4 model achieves a 12.8% enhancement in mAP for the pedestrian class in Sequence 1.

These seven works constitute the memorandum of this Ph.D. thesis, and the author presents the results of these years of research.

Chapter 1

Introduction

We should not let our fears hold us back
from pursuing our hopes.
John F. Kennedy.

Abstract: This first chapter serves as a general introduction to this thesis's context, motivation, problem, and methodology. A structure of the document is also provided to make it easier to follow.

1.1 Context

The rise of technology has been a revolution that has led to a notorious change in society. Among other things, the **accessibility** and **reduced cost** of devices and various services have promoted this, changing how we live and work. A few years ago, simple personal computers or Internet access supposed a high cost that only a few could afford. Today, reducing the cost of these devices has made them more accessible. As a result, the **amount of information** and content available has **increased dramatically**.

The reduction in the cost of these systems is not directly related to computers. It goes beyond them. This fact also affects the rest of the electronic devices, where they have been improved over the years. Among these devices, for example, there are systems focused on collecting images and videos, such as video surveillance cameras. The information collected by these devices is easily accessible through several platforms on the web. Today, many systems are specialized in different tasks related to the collection of multimedia content. The most common are 2D cameras, which collect photos or video sequences for further processing. Some examples could be the camera that comes with smartphones or video surveillance systems used in public environments, where they are usually placed at high points for surveillance or traffic control, among others. Figure 1.1 shows some types of cameras and devices that allow collecting this information. There are also in-vehicle systems, called dashcams, that help drivers by filming the road from their point of view.

Figure 1.1: Examples of different types of video surveillance camera devices using an image from Elharrouss et al. (2021). (a) Fixed camera. (b) PTZ external camera. (c) PTZ (Pan-Tilt-Zoom) internal camera. (d) Bullet camera. (e) Dome camera. (f) Turret camera.

The data collected by these devices can be for personal or professional purposes. The latter group collects data for private use, either for a company or public organizations and entities, such as the government, with specific purposes. Examples include installing cameras and video surveillance systems on city streets to control crowds in a particular area or roadways to monitor traffic and identify potential hazards.

With the rise of these types of systems, there is a substantial increase in data collection, creating several **challenges** and issues that need to be addressed. In many cases, such as the personal use of these devices, the information collected is dispensable and used for entertainment purposes only. However, there are other areas where the analysis of the collected data is essential. Otherwise, collecting this information would be useless. According to road management and monitoring, **video surveillance systems** capture images 24 hours a day. It is a challenge to manage and analyze the images these systems capture. A single recording that covers an entire day requires evaluating several designated people. Providing **new methods** and advances to analyze these sequences optimally and efficiently is necessary.

In the last few years, many advances have been developed in the field of CV, making it possible to process images and videos according to established requirements. On the other hand, thanks to the increase of this data, the field of DL has significantly been influenced, leading to the development of models that learn from this information. All this has resulted in a set of models for both detection and segmentation that allow for identifying a wide range of elements. However, they still have **shortcomings** when it comes to identifying small elements. Their

mAP (Mean Average Precision) decreases significantly when the size of these elements is small. All this is partly due to the intrinsic behavior of the model itself. When performing the inference, it reduces the number of pixels of the image given as input. As a result, the elements composed by a low rate of pixels are lost in the processing. Thanks to the proposals presented in this thesis, it has been possible to improve the mAP for a wide range of object detection and segmentation models without the need to modify the internal structure of the model. DA (Data Augmentation) and re-referencing techniques are applied, providing a methodology to improve the identification of these types of elements in different sequences, mainly focused on road environments.

1.2 Problem and Goal

In the context of video sequence analysis with CV, this thesis aims to improve the identification of small objects in video sequences using techniques from DL and SR.

A small object is defined as an object consisting of a few pixels. Depending on the evaluated context, there are many types of these elements. This thesis focuses mainly on elements that appear in pedestrian areas and streets, such as pedestrians or different types of vehicles.

The main approach stated in this thesis has been the development of new **methodologies** to **improve** the **identification** of **small objects** and, in general, the mAP through the use of different models oriented to the detection and segmentation of objects, avoiding modifying the architecture of these. Many models have been developed for detection and segmentation as part of CNN. These models achieve good results in identifying a large number of elements. However, the **accuracy decreases** significantly as the size of these elements is reduced. The approach proposed in this thesis has mainly consisted of working on the robustness of different pre-trained detection and segmentation models to identify new small objects and improve the accuracy of the other elements, especially in traffic scenes. These methods are mainly based on increasing the number of pixels that constitute the elements contained in the image using SR techniques and different strategies for selecting the areas to apply this procedure to subsequently re-infer it.

A second line of work is using object detection networks for anomaly detection. This approach is based on applying techniques that improve the detection of elements without modifying or retraining the model to compute the common position of elements that belong to the same class. Then, in areas where an element does not normally pass, anomalous behavior is detected.

The objectives achieved in this thesis are detailed below:

1. Enhancement in Small Object Detection:

- Development of methodologies to enhance the identification of small objects by utilizing models oriented towards object detection.
- Application of data augmentation techniques, such as Super-Resolution (SR), to increase the number of pixels constituting these objects and facilitate their detection.
- Utilization of pre-trained Convolutional Neural Networks (CNNs) for object detection and methods provided for Fine-Tuning through unsupervised training.

2. Improvement in Small Object Segmentation:

- Development of methodologies to improve the segmentation of small objects in video sequences.
- Application of Super-Resolution (SR) techniques and area selection strategies to enhance the robustness of pre-trained detection and segmentation models.

3. Anomaly Detection in Small Elements:

- Employment of object detection networks for anomaly identification.
- Application of techniques to improve the detection of elements without modifying or retraining the model.
- Detection of anomalous behaviours in areas where certain elements are generally not expected.

1.3 Methodology

The execution time of a method can be a critical factor, depending on the area in which it is used. Therefore, it is necessary to provide solutions where the required complexity is as low as possible to solve problems that need to be implemented in efficient environments. However, in some domains, it is more important that the proposed method accurately identifies most elements in the scene. Therefore, the application domain will determine the method's quality and accuracy.

Some general methodological principles have been followed to give an adequate solution to the problems proposed in this thesis:

- **Scientific Method**: This method is the research foundation by ensuring reproducibility and refutability. In other words, all experiments performed on the same dataset must produce the same results to be reproducible by other researchers. It also assumes that any scientific hypothesis can be refuted. A series of well-defined steps characterize this method:

  1. Observation: Stage in which the real problems are analyzed and studied, along with the existing models, to provide new and more robust methodologies to solve them.
  2. Hypothesis: The approach of an improvement in identifying small elements using techniques that support DL models is determined as a hypothesis.
  3. Experimentation: A set of experiments are performed to test the proposed methodologies in real scenarios.
  4. Analysis: The performance of the proposed methodology is compared with other existing solutions.
  5. Conclusions: According to the previous analysis, a set of conclusions is reached to confirm or reject the proposed hypothesis.

- **Iterative and incremental development methodology**: The project is being developed based on progressively implemented functionalities and improvements. Results have guided the implementation of new advancements to improve the proposed methodology and address problems observed during the development process.
- **Implementation methodology**: The methodologies have been designed to be modular, maintainable, and extensible to make development and usage easier. In addition, each of them is adequately documented.
- **External Evaluation Criteria**: The methodology uses performance measures commonly applied in this field, such as the difference between the accuracy of the model's results and the GT (Ground Truth). Furthermore, widely used datasets have been selected to enhance comprehensibility and enable comparison with other research.
- **Comparison with other methods**: In cases where feasible, a comparison has been made between the outcomes of this methodology and other advancements for a similar experiment. With this objective in mind, the intention is to outline the advantages and disadvantages of the proposal in comparison to alternative options, ultimately establishing its originality and substantial contribution.

1.4 Structure on this Thesis

This Ph.D. thesis can be divided into three differentiated sections, not including the appendices. The first section constitutes Chapter 2, which consists of the theoretical background of the thesis. Therefore, it details the background and the state of the art, which was used as a starting point for the research process, supporting the works included in the thesis. This research constitutes the second part and covers the Chapters 3 to 9. The third and last part presents the conclusions obtained concerning the research process conducted and the most promising potential lines of work to be performed in the future to continue the development of the different methodologies presented.

In terms of theoretical foundations, the basics of each component that composes CNN models is described in detail to provide a comprehensive understanding. Next, the fundamentals of DA (Data Augmentation) networks are discussed, focusing on those related to SR. The principles of some of the most important models are described, along with an overview of the solutions that can be provided using these models in different areas. Once the models focused on identification and SR are generally established, the problem of detecting small objects is described in detail. The definition of a small element and a description of the problem in this area are established. This point is one of the most extensive parts of this thesis. Then, the problem of detecting anomalies and threats in video sequences is described in detail.

The second part contains the works that support this thesis. Each of these works constitutes a specific chapter.

The first of the nine papers that compose this thesis is presented in chapter 3. Entitled *Improved detection of small objects in road network sequences using CNN and super resolution*, it was published in 2021 in the journal *Expert Systems*,

which was ranked Q2 (37/110) of the JCR (Journal Citation Report) ranking in the *Computer Science, Theory & Methods* category. This work focuses on **improving** the **identification** of **small objects** by using different pre-trained object detection models and techniques that increase the image resolution. For this purpose, a methodology is proposed based on the **re-inference** of the previously **super-resolved** areas where some objects have been located. For this purpose, several experiments using different models are presented to determine the improvement provided by the proposed methodology.

The chapter 4 and the second paper presented is entitled as *Enhanced Image Segmentation by a Novel Test Time Augmentation and Super-Resolution* and was presented at the 9th IWINAC (International Work-Conference on the Interplay Between Natural and Artificial Computation) in 2022. The paper deals with improving the detection of **small objects** in road sequences by applying different **object segmentation models**.

The third of the presented papers constitutes the 5th chapter of this thesis and is entitled *Optimized instance segmentation by super-resolution and maximal clique generation*. It was submitted in 2023 to the journal ICAE (Integrated Computer-Aided Engineering), with the ranking Q1 in the JCR (Journal Citation Report) ranking in Computer Science, Interdisciplinary Applications (22/110). This work follows studies related to improving the segmentation of small objects. For this purpose, a **heuristic** is proposed to **optimize the execution times**, allowing its application in those environments where time is a restrictive requirement.

The fourth paper belongs to the chapter 6 titled *Automatic labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks* published in the journal PRL (Pattern Recognition Letters) in 2023. The paper presents a methodology for **fine-tuning** an unsupervised object detection model with shortcomings in identifying scene elements, removing the need for a previously annotated dataset.

The fifth of the presented papers constitutes chapter 7 of this thesis, named *Minimal Optimal Region Generation for Enhanced Object Detection in Aerial Images using Super-Resolution and Convolutional Neural Networks*. It was presented in 2023 at the 17th IWANN (International Work Conference on Artificial Neural Networks). The work involves determining areas where most objects in a video sequence are gathered. Based on these calculated zones, a set of **minimal regions** is **optimally** determined to be used for re-inferring after applying DA (Data Augmentation).

Chapter 8 is the sixth paper called *Object detection in traffic videos: an optimized approach using super-resolution and maximal clique algorithm*. The article was published in the journal *Neural Computing and Applications* in 2023 and was ranked Q2 (41/145) in the AI (Artificial Intelligence) category of the JCR classification. This work combines **optimization techniques** with SR techniques to **reduce** the number of **areas** to be inferred using an object detection model, thus significantly improving the execution time.

The last work and chapter 9 and the seventh paper is entitled *Small-Scale Urban Object Anomaly Detection Using Convolutional Neural Networks with Probability Estimation*. The paper was published in the journal *SENSORS* in 2023, and it was ranked Q2 (100/275) in the Engineering, Electrical & Electronic category of the JCR classification. This paper presents a methodology for **anomaly detection** in **synthetic video sequences**. It deals with identifying small objects

and improves the performance of object recognition models without changing their internal structure.

The third and last part of this thesis summarizes the conclusions of all the work presented and identifies the most promising lines for future work.

Chapter 2

Fundamentals and State of the Art

All men dream; but not equally. Those who dream by night in the dusty recesses of their minds Awake to find that it was vanity; But the dreamers of day are dangerous men. That they may act their dreams with open eyes to make it possible.

Seven Pillars of Wisdom: A Triumph,
T.E. Lawrence.

Abstract:

This chapter explains the concepts, theory, and related work behind this PhD thesis. It includes the basis of CNN, fundamentals of SR, small object identification, and anomaly detection using DL techniques.

2.1 Fundamentals of Neural Networks and Deep Learning

Neural networks have revolutionized and provide solutions to **complex challenges** based on image processing and multidimensional data. Therefore, it has become a fundamental paradigm in DL and AI. With the significant increase in data generation, neural networks have made it possible to provide **answers** to complex problems in several areas, such as CV.

Their structure is similar to the capacity and structure of the human brain, where biological neurons communicate and collaborate to perform a series of specific tasks related to cognition and perception. Figure 2.1 illustrates the similarities between both of them.

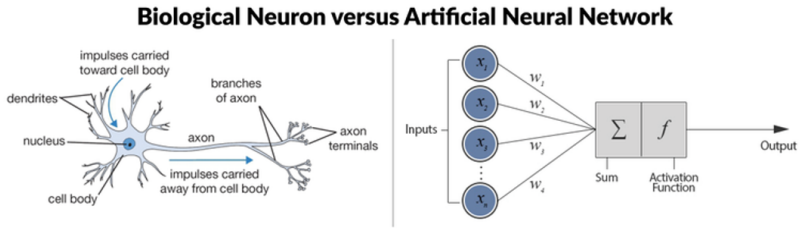

Figure 2.1: Comparison between natural and artificial neuron schemes using an image from Moumene and Ouelaa (2022).

CNN (Convolutional Neural Networks) maintain this notion based on the interconnection of neurons that transmit and transform information through the composition of a series of **layers**, which constitute the processing units called **artificial neurons**. The behavior of a neuron is roughly based on the reception of multiple weighted inputs. The sum and application of a specific activation function produce an output. The **organization** and **interconnection** of these neurons make it possible to learn patterns and abstract representations from the data given as input.

This thesis comprises a series of works focused on using this model type. Therefore, the basic principles and foundations that support the understanding and use of these networks are established below.

2.1.1 Artificial Neurons

The neuron is the most essential component of a CNN. It performs a weighted and nonlinear operation on the input data. It allows the network to learn by detecting patterns and features in the data. Therefore, it can be defined as a mathematical function that takes one or more values as input and outputs a single numerical value. It is determined by the formula below:

$$y = f\left(\sum_{i=1}^{n} x_i w_i + b\right) \tag{2.1}$$

The first step is to compute the sum $\sum_{i=1}^{n} x_i w_i$, where x_i represents each of the n inputs and their synaptic weights w_i, also known as activation values. Since a **neural network** comprises a series of **neurons** in different **layers**, the value x_i can represent either a value associated with the input data or the output of another neuron in the layer immediately preceding it. On the other hand, the weights of w_i determine the influence of this input on the output. The weighted input of this neuron is obtained by summation.

The bias, denoted by the variable b, is a constant value added to the neuron's weighted input before applying the activation function. It is intended to allow the neuron to adjust the base activation.

The result of this sum is the input to the activation function f, also known as the **transfer function**. This function introduces nonlinearity in the neuron output and is fundamental in determining complex relationships in the data. A

wide variety of activation functions are explained in more detail in the subsection 2.1.2.2. It results in the output of the neuron y, a single computed value.

2.1.2 Layers

CNN are composed of a series of **layers** that work together to **extract relevant features** and patterns from the input information. Each layer plays a fundamental and specific role in this process. Therefore, a detailed description of each layer is given below.

2.1.2.1 Convolutional Layers

This layer is the starting point of a CNN. According to image processing, this layer takes as input an image I represented by a three-dimensional matrix $H \times W \times C$ representing the height, width, and RGB (Red-Green-Blue) color channels, respectively. The parameters composing these layers consist of a set of K filters, also called **kernels**, which are mostly square in shape, with a fixed width and length. The kernels cover the full dimensionality of the channels that compose the input image and act as a feature detector to search for specific patterns in the input data.

The convolution operation is performed by **sliding** these filters over the input, computing the scalar product between the filter and the corresponding input region. All this results in a feature map that highlights the presence of certain **patterns**. Thus, given an input I with dimensions $H \times W \times C$ and a filter K with dimensions $F \times F \times C$, the convolution operation at a specific position (i, j) of the feature map will be computed as follows:

$$\text{Conv}(I, K)(i, j) = \sum_{a=1}^{F} \sum_{b=1}^{F} \sum_{c=1}^{C} I(i + a - 1, j + b - 1, c) \cdot K(a, b, c) \tag{2.2}$$

Where I is the input data, H and W are its height and width, respectively, K is the applied filter, C is the depth or number of channels, and F is the size of the filter. This process is performed for all input data positions and each convolutional layer filter. The result is a set of feature maps with relevant information for subsequent tasks.

2.1.2.2 Activation Layers

After the convolution operation, a specific **nonlinear activation function** is applied to model complex nonlinear relationships in the data. As a result, the network can capture much more abstract features and patterns. Technically, it is not considered as such layers since no parameters or weights are learned within the layer.

These types can make use of a variety of nonlinear activation functions. The most common ones are listed below:

- **Linear Function (LINEAR):** It is one of the simplest functions because it does not introduce nonlinearity. This function is used in cases where an

output proportional to the input is desired. It is rarely used in hidden layers. It is mainly used in the output layer for linear or regression models.

$$\text{Linear}(x) = x \tag{2.3}$$

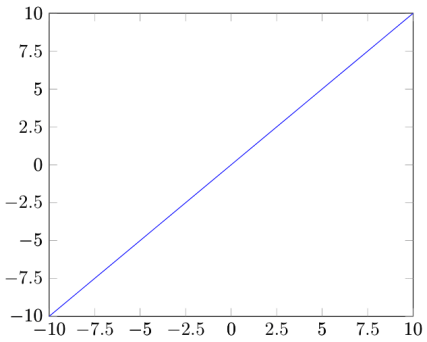

Figure 2.2: Linear function plot.

- **Rectified Linear Unit Function (ReLU):** It is one of the most commonly used activation functions. Its simplicity and the introduction of **non-linearity** in the network characterize it. It efficiently solves gradient fading problems in deep layers by converting negative values to zero and leaving positive values unchanged. One of the problems is that if a neuron's values are always negative during training, it may be prevented from updating its weights, preventing the network from learning.

$$\text{ReLU}(x) = \max(0, x) \tag{2.4}$$

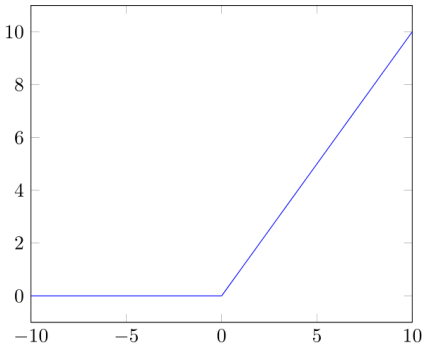

Figure 2.3: ReLU function plot.

- **Leaky ReLU:** It is a **variant** of RELU (Rectified Linear Unit Function) that allows a small value for negative values instead of transforming them directly to zero. This solves the problem of the learning contribution of a neuron with always negative values.

$$\text{Leaky ReLU}(x) = \begin{cases} x, \text{si } x > 0 \\ \alpha \cdot x, \text{si } \ x \leq 0 \end{cases} \tag{2.5}$$

α represents a small and negative value. This is usually set to 0.01.

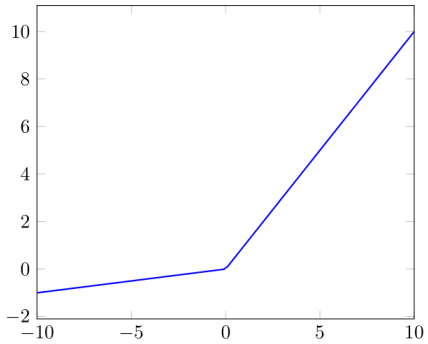

Figure 2.4: Leaky ReLU function plot.

- **Sigmoidal Function (Logistic):** This function maps the values in a range between 0 and 1. The resulting plot is in the form of S. It is mainly used in **hidden layers** to introduce nonlinearity and to control the outputs. The problem of gradient fading can occur because the derivatives are small for large or small values.

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

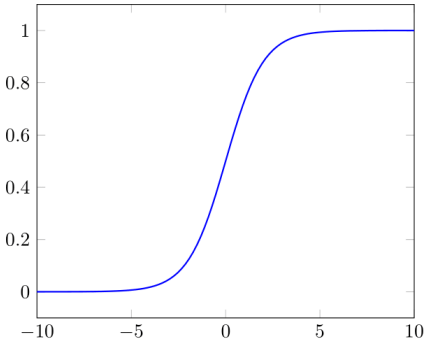

Figure 2.5: Sigmoidal function plot.

- **Hyperbolic Tangent Function (Tanh):** Like the sigmoid function, it has an S shape. However, the values are mapped in a range between -1 and 1. One of the differences to the sigmoid is that it has a **symmetric range** around the origin, which allows to mitigate the problem of data centralization as much as possible.

$$\text{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.7}$$

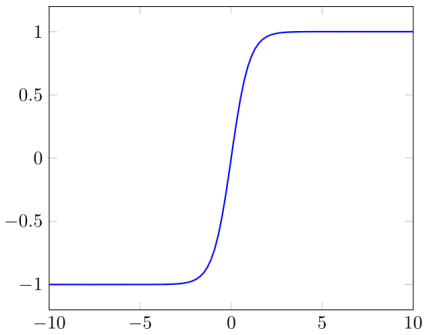

Figure 2.6: Tanh function plot.

- **SoftMax Function (SOFTMAX):** It is applied to a set of values and returns a **probability distribution**. It is often used in the **output layer** in connection with classification problems, where probabilities are assigned to each class. According to a set of values x_1, x_2, ..., x_n:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2.8}$$

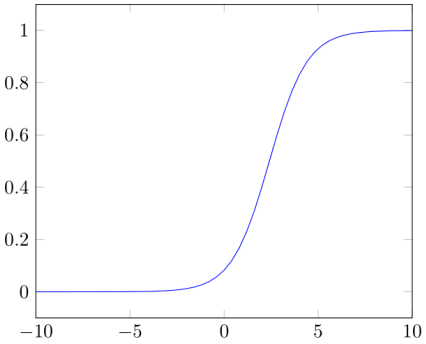

Figure 2.7: SOFTMAX function plot.

2.1.2.3 Pooling Layers

Pooling layers reduce the **spatial dimensionality** of the **features**, reducing the number of parameters in the network by controlling overfitting. A very common technique is known as max-pooling, where the feature matrix is divided into several regions of size P *times* P, and the maximum value of each of the regions is selected.

$$\text{MaxPool}(M)(i,j) = \max_{a=1}^{P} \max_{b=1}^{P} M(i \cdot P + a, j \cdot P + b) \tag{2.9}$$

M is the feature map in which the max-pooling MaxPool(M) is applied at position (i, j), where these are the coordinates in the feature map. P defines the size of the filter by specifying the maximum value within each block in the map M.

2.1.2.4 Fully Conected Layers

Fully connected layers are a fundamental component of neural networks. In these layers, each neuron is **connected** to **all neurons** in the **previous and next layers**, allowing for comprehensive information processing. Each neuron sums its weighted inputs and applies an activation function, allowing it to learn complex relationships in the data. These layers are critical for classification and regression tasks and play a fundamental role in DNN (Deep Neural Networks).

2.1.2.5 Batch Normalization Layers

This type of layer is an **important addition** to the neural network architecture and was introduced by Ioffe and Szegedy (2015) to **improve** training stability and performance. It normalizes a network's intermediate activations by adjusting their mean and standard deviation. This makes it possible to mitigate problems related to fading, which occurs when the normalized activations become too small when going deeper into the network, or gradient explosion where activations become too large. It also speeds up training convergence, reducing the number of epochs or stages required. A drawback is that it can slow down the training time due to batch statistics computation and normalization, even if it requires a small number of epochs.

$$\mathrm{BN}(x) = \gamma \frac{x - \mu}{\sigma} + \beta \tag{2.10}$$

X is the input to the layer, μ is the mean of the mini-batch, σ is the standard deviation, and γ and β are the parameters that allow normalization adjustment.

2.1.2.6 Dropout Layers

They were introduced as a technique to **prevent overfitting** and are accomplished by **randomly deactivating** a set of neurons with a certain probability Srivastava et al. (2014). This prevents the network from becoming dependent and improves generalization. The process can be expressed simply as:

$$\mathrm{Dropout}(x) = x \cdot \mathrm{mask} \tag{2.11}$$

Where x is the input to the layer, and mask is a binary operation randomly applied to the neurons. In the evaluation phase, these layers are not applied. Instead, the values are set based on the inverse of the probability to maintain the expected value.

2.1.3 Loss Function

Loss functions are a fundamental part of CNN. They are responsible for **measuring the discrepancies** between the model predictions and the actual values, guiding the optimization process by adjusting the network weights, and **improving** the **generalization** of previously unseen data. There are many loss functions. The most common ones are listed below:

- **Mean Squared Error (MSE):** Its objective is based on **minimizing** the sum of squares of the differences between the predictions given by the model, set as y_{pred}, and the actual values y_{actual} for each training sample.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_{\text{pred},i} - y_{\text{real},i})^2 \tag{2.12}$$

  N is the number of training samples. It is used when desired to **penalize** significant discrepancies between predictions and actual values significantly.

- **Mean Absolute Error (MAE):** It is responsible for measuring the average difference between the absolute value of the predictions y_{pred} and the actual values y_{actual} for each training sample.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_{\text{pred},i} - y_{\text{real},i}| \tag{2.13}$$

  It is usually used when discrepancies are **smaller**.

- **Binary Cross Entropy (BCE):** Intended for use in **binary classification problems**. It measures the differences between the predicted probabilities p_{pred} and the actual labels for each class p_{actual}.

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(p_{\text{real},i}\cdot\log(p_{\text{pred},i}) + (1-p_{\text{real},i})\cdot\log(1-p_{\text{pred},i})\right) \tag{2.14}$$

  Where N is the number of training samples.

- **Categorical Cross Entropy (CCE):** This function is used in **multi-class classification** problems. It is similar to the BCE (Binary Cross Entropy). However, it deals with more than two classes. It is expressed as:

$$\text{CCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{\text{real},ij}\cdot\log(y_{\text{pred},ij}) \tag{2.15}$$

  N is the number of training samples, and C is the number of classes. $y_{realpred}, ij$ is a variable that takes the value 1 if sample i belongs to class j or 0 otherwise.

As has been noted, each loss function has several advantages and disadvantages. Therefore, to optimize the model's performance, it is required to choose which one to use according to the scope of the problem.

2.1.4 Training Strategies

Several strategies have been developed to improve the performance of CNN. This process is not simply based on parameter tuning. It involves a series of strategies to improve the model's accuracy and generalization.

2.1.4.1 Training from scratch

This process starts with a CNN architecture initialized with **random weights**. During the training process, these weights are **adjusted** to minimize the loss function from the training data set. This updating of the weights is done through several optimization techniques and algorithms. One of the most notable is SGD (Stochastic Gradient Descent). The updating of these weights in a given iteration is expressed as

$$W_{t+1} = W_t - \alpha \nabla L(W_t, X_{\text{batch}}, Y_{\text{batch}}) \tag{2.16}$$

Where W_t are the weights of iteration t, α is the learning rate. ∇L is the gradient descent with the loss function L and the weights W_t. X_{batch} and Y_{batch} are the data and labels of the training batch.

2.1.4.2 Fine-Tuning

This approach starts with a **pre-trained** network on a generic dataset. It is used when the network performance needs to be **extended** to new data. For this purpose, the initial layers are kept **frozen** or at a very low adjustment. The final layers, corresponding to the fully connected layers, are adjusted to adapt to the new task. The weights are updated similarly to training from scratch. It has the peculiarity that the initial layers can be kept fixed at $\nabla L = 0$ or at a small rate. This technique has many advantages, such as using the network's prior knowledge or saving time and resources.

2.1.4.3 Transfer Learning

This technique is similar to fine-tuning, except it may involve adjusting some **intermediate layers** rather than just the **final ones**. This is done by adjusting only part of the network to adapt to the new task while retaining the features learned in previous layers.

2.2 Fundamentals of Super-Resolution

The increase and improvement of **image quality** is one of the areas of research and development in CV. The main objective of this type of technique is to produce an image with a **higher resolution** and **level of detail** compared to the original image. Several fields can benefit from this technique, such as improving security through images captured by video surveillance systems or a more accurate diagnosis in the medical environment.

The systems designed for capturing sequences have been improved due to the demand for high-quality images. However, applying several techniques that support these systems to provide better image quality is latent. This line of research has recently undergone a remarkable **transformation**. From the beginning, quality improvement strategies were limited to **preserving fine details** and **recomposing visual information** not present in the original image. New approaches have been explored using DL to obtain **high-quality images** based on their low resolution.

This thesis presents a series of works that apply different models focused on SR as a previous step to improve the quality of the images. Therefore, this section details the classical techniques applied to quality improvement, SR basics, and the most prominent models.

2.2.1 Classical Techniques for Quality Improvement

The first approach was based on classical techniques such as **filtering** or **interpolation** in image quality improvement. The most commonly used techniques are described below:

- Bilinear Interpolation: This technique is one of the simplest and most commonly used methods to increase the resolution of an image. Let I_{LR} be a low-resolution input image. A high-resolution version I_{HR} is desired by applying bilinear interpolation. Therefore, the value of each pixel in the image I_{HR} is computed as a **weighted combination of the neighboring pixels** in I_{LR}. Given a pixel with coordinates (x, y) in I_{HR}, its value in the low-resolution image is calculated as:

$$I_{HR}(x,y) = \sum_{i=-1}^{1} \sum_{j=-1}^{1} w_{i,j} \cdot I_{LR}(x+i, y+j) \tag{2.17}$$

  $W_{i,j}$ determines the interpolation coefficients that define the weighting of neighboring pixels.

  The main feature of this method is its simplicity of application, which, in many cases, gives acceptable results. However, in those conditions where it is necessary to obtain much sharper high-resolution images, bilinear interpolation does not allow significant quality improvement based on the low-resolution image.

- Filtering: Filtering is another classic approach to quality improvement. It is based on applying **filtering operators** to the image, allowing the emphasis or removal of specific image features. This method includes a variety of filters, such as Gaussian smoothing, designed to remove noise and irregularities in the image. If I_{LR} is a low-resolution image and G is the Gaussian kernel, the filtering process can be expressed as:

$$I_{HR}(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} G(i,j) \cdot I_{LR}(x+i, y+j) \tag{2.18}$$

  K determines the size of the kernel. On the other hand, $G(i, j)$ determines the value of the Gaussian kernel at position (i, j).

  Again, there are several shortcomings. It can also lead to a **loss** of **fine detail** and cause the edges of elements to blur.

Although these techniques have been widely used, there are limitations as mentioned above. In response to these shortcomings, it is necessary to develop new and much more advanced approaches to take advantage of DL and effectively increase the quality and resolution of images. SR (Super-resolution) is the answer

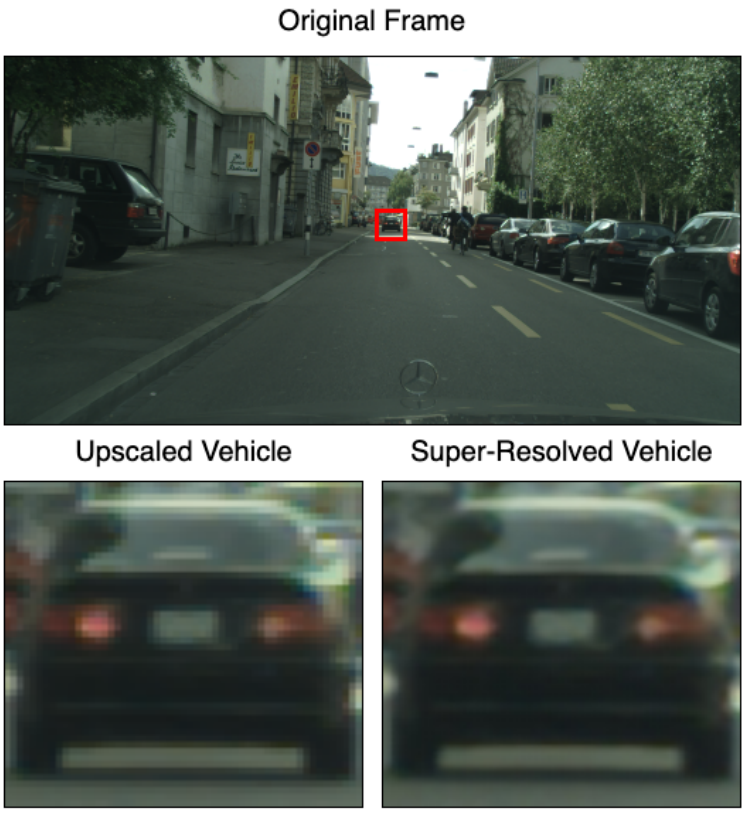

Figure 2.8: Comparison between the upsampling and super-resolution using an image from the CityScapes Dataset Cordts et al. (2016).

to this problem. Figure 2.8 shows the differences between the upsampling and super-resolution.

2.2.2 Super-Resolution Fundamentals

SR is a technique that aims to increase the resolution of an image of low quality or resolution given as input. Once the image is processed, a **higher** definition and **level of detail** is obtained. It differs from the previous classical techniques because it uses **contextual information** and **intrinsic relationships** present in the data to infer details and produce a super-resolved image.

This is done by **analyzing patterns and features** in the low-resolution image. The goal is to improve resolution by searching for patterns and features in high-resolution images that can be applied to low-resolution images. Images contain diverse information, including repeating patterns, textures, and visual structures resulting from real-world features such as sharp edges between objects. By exploiting these spatial relationships, SR techniques learn the features associated with patterns and structures.

Let I_{LR} and I_{HR} be the low and high-resolution images, respectively. The

pixels in I_{LR} are represented as $I_{LR}(x_i, y_i)$, under the coordinates (x_i, y_i). Using a function R that allows to map the pixels of I_{LR} to I_{HR}, the spatial relationship can be expressed as follows:

$$I_{HR}(x_j, y_j) = R(I_{LR}(x_i, y_i)) \tag{2.19}$$

Where (x_j, y_j) are the coordinates associated with the high-resolution image I_{HR}. By modeling these relationships, SR algorithms and techniques can learn how to **generate realistic details** from low-quality images, such as edge enhancement or textures.

Regularization is also a fundamental component of these algorithms since restoring detail must be realistic and avoid introducing artifacts or noise. Through regularization, unrealistic solutions are penalized by a loss function used to train the model.

2.2.3 Models

SR has increased the number of models aimed at improving image quality and resolution in a wide range of areas. These models can be divided into two main groups based on CNN or GAN (Generative Adversarial Networks). Each of these models approaches the challenge of quality improvement differently. The desired effectiveness in the application context determines their selection.

2.2.3.1 Models Based on Convolutional Neural Networks

Adopting architectures based on CNN has led to several models aimed at SR. They assume that the low-resolution images I_{LR} contain **enough information** to perform texture and detail restoration in the SR image I_{HR}. This is done by the model identifying and learning correlations between the two resolutions in the training data. Let F be the selected SR model:

$$I_{HR} = F(I_{LR}) \tag{2.20}$$

Where I_{HR} is the high-resolution image obtained by the mapping relation previously learned by the model. Several models like ESPCN (Efficient Sub-Pixel Convolutional Neural Network) Shi et al. (2016), EDSR (Enhanced Deep Super-resolution Network) Lim et al. (2017), LAPSRN (Laplacian Pyramid Super-Resolution Network) Lai et al. (2017), or FSRCNN (Fast Super-Resolution Convolutional Neural Network) Dong et al. (2016b) are presented in this group. One of the models that stands out in this category is the last.

FSRCNN (Fast Super-Resolution Convolutional Neural Network) model Dong et al. (2016b) is characterized mainly by its computational efficiency and ability to process low-resolution images **quickly**. The model is based on the SRCNN (Super-Resolution Convolutional Neural Network) Dong et al. (2016a), which proposes an hourglass-shaped neural network structure. The most expensive task in super-resolution models is mainly feature mapping. In this model, the input size is first reduced (shrinking layer) to reduce the complexity during the mapping phase. The next step is expanding the input size (expansion layer) to increase the dimension of the features. This is achieved while maintaining good detail in the output, resulting in an efficient and fast model.

The feature extraction and reconstruction phases can be described as follows:

$$I_{feat} = \phi(W_{feat} * I_{LR}) \tag{2.21}$$

$$I_{HR} = \phi(W_{up} * I_{feat}) \tag{2.22}$$

I_{feat} represents the low-resolution image features I_{LR}, while W_{feat} and W_{up} are the weights of the convolutional feature extraction and reconstruction layers, respectively. ϕ represents a specific activation function.

This convolutional neural network model has many advantages, such as automatically learning patterns and features from images. Without predefined features, the model can recognize these intrinsic relationships between low- and high-resolution images. On the other hand, its deep architecture allows it to capture features with multiple levels of abstraction, thus recovering finer details in the processed images.

The choice of a particular model is determined by several requirements, such as the computational complexity or the visual quality obtained after applying the model. The FSRCNN (Fast Super-Resolution Convolutional Neural Network) model is suitable in situations where resource optimization is essential, which is one of the reasons why it was selected. Figure 2.9 shows the network structures of the SRCNN (Super-Resolution Convolutional Neural Network) and FSRCNN (Fast Super-Resolution Convolutional Neural Network) model.

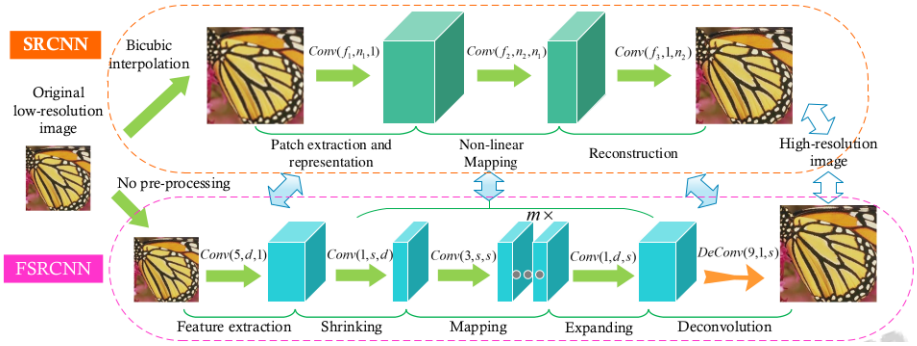

Figure 2.9: Comparison between the SRCNN and FSRCNN model network structures using an image from Dong et al. (2016b).

2.2.3.2 Models Based on Generative Adversial Networks

The incorporation of GAN for SR has led to a significant improvement in this area. GAN (Generative Adversarial Networks) are a type of model based on competition between two networks. A G network is known as the **generator**, and a D network is known as the **discriminator**. In brief, the operation is as follows. The generator tries to **generate** several high-resolution, realistic images, while the discriminator tries to **distinguish** which images were generated and which were real. This competition between the two networks results in continuous improvement, as G learns how to generate increasingly convincing high-resolution images while D increases and improves its ability to distinguish between the real and the generated.

This category includes models such as SRGAN (Super-Resolution Generative Adversarial Networks) Ledig et al. (2017), ESRGAN (Enhanced Super-Resolution Generative Adversarial Networks) Wang et al. (2019), or REALESRGAN (Real Enhanced Super Resolution Generative Adversarial Network) Wang et al. (2021), among others. This last model, the REALESRGAN (Real Enhanced Super Resolution Generative Adversarial Network), is characterized mainly by the quality of the generated images. It combines the power of GAN with image processing techniques to generate high-resolution images with a realistic visual appearance that maintains fidelity compared to low-resolution images. This generation process can be represented as follows.

$$I_{HR} = G(I_{LR}) + E(I_{LR}) \tag{2.23}$$

The super-resolved image I_{HR} is generated by G, which produces an initial high-resolution version, and E, which corresponds to the detail enhancement stage responsible for refining the generated image by adding high-frequency details to improve its perceptual quality and realism.

In addition, the architecture of the REALESRGAN (Real Enhanced Super Resolution Generative Adversarial Network) model may include attention mechanisms that allow the model to focus on specific regions of the image during its generation, improving the ability to capture and highlight details. Figure 2.10 shows the architecture of this model.

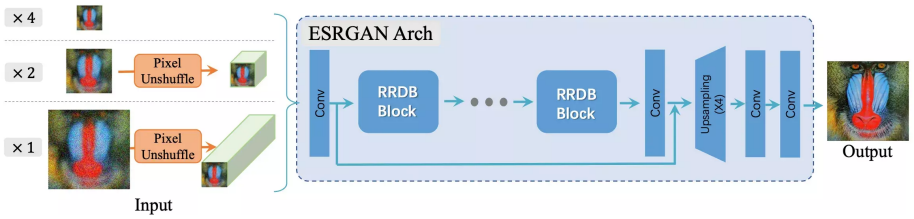

Figure 2.10: Real-ESRGAN architecture, which adopts the same generator network as the ESRGAN model using an image from Wang et al. (2021).

2.3 Detection and Segmentation of Small Objects

Accurately recognizing and detecting objects in images and video sequences is a critical challenge in various applications, such as medicine or video surveillance control and monitoring. One of the most complex challenges in this area is the identification of small objects with limited dimensions. These objects are often extremely small, requiring new methods to identify them effectively. CNN have been a key element in object recognition. However, the mAP (Mean Average Precision) of CNN (Convolutional Neural Networks) is significantly lower for identifying these elements.

This section discusses the fundamentals based on object recognition and their respective detection and identification in detail to provide a more comprehensive understanding. This thesis is supported by a series of works based on improving the accuracy, mainly of small objects, throughCNN (Convolutional Neural Networks).

Therefore, the problem inherent in detecting small objects, possible solutions, the challenges that need to be addressed, and the applications and domains that can benefit from solving this problem are presented in detail. This exploration culminates in specifying current solutions and detailing proposals contributing to this field.

2.3.1 Object Recognition

Object recognition, also known as image classification, is a fundamental task in CV (Computer Vision) and AI (Artificial Intelligence). It is mainly based on the correct and automatic identification and labeling of objects and visual patterns in an image or video sequence. Therefore, this task is essential in many areas. One example is autonomous driving, which identifies traffic signs. Figure 2.11 shows an example of an image and the expected output after performing image classification.

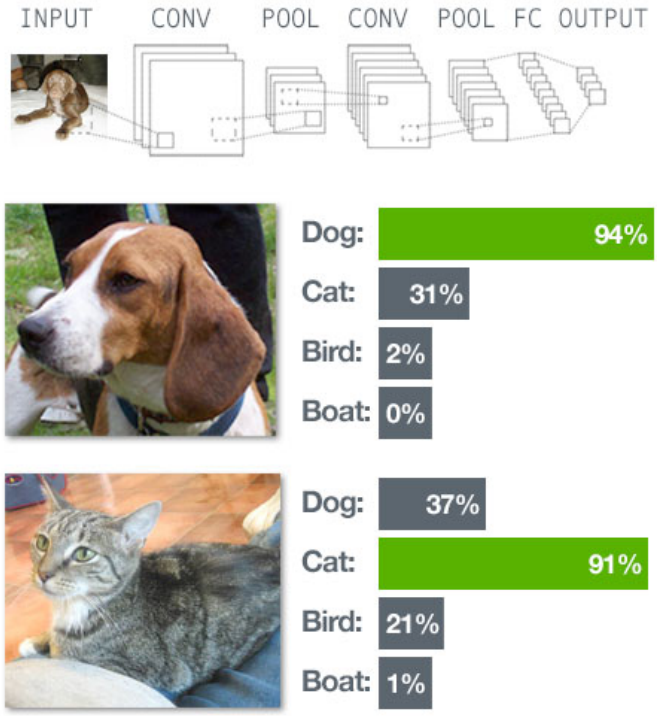

Figure 2.11: Example of an image and the expected output after performing image classification using an image from Duraisamy et al. (2022).

Initially, object recognition relied on several traditional approaches that involved **manually extracting** visual features such as edges, colors, or textures to identify the elements present in the images. Bay et al. (2008) presents a technique based on manually creating **feature detection filters** and multiplying images with them. Each filter was designed to detect specific features. The combination resulted in the classification. One of the best-known feature descriptors is the HOG (Histogram of Oriented Gradients) Dalal and Triggs (2005). This technique

has been used to determine the shape and texture of an object in an image. It calculates the magnitude and orientation of the image gradients in several local regions. These orientations are then used to construct a histogram. Although these techniques were successful in specific applications and under controlled conditions, they had **several limitations** when the **variability** was **complex**, i.e., they had to deal with aspects such as changes in illumination, occlusion, scale variability, or orientation of the elements, requiring more robust alternatives.

Advances in CNN (Convolutional Neural Networks) represented a revolution for this type of task. These networks proved very effective, allowing the automatic learning of discriminative features from the input data. As a result, the need for manual feature extraction was eliminated. This opened up new lines of research for the development of techniques focused on accuracy and robustness in the field of object recognition.

One of the first CNN (Convolutional Neural Networks) was developed by Lecun et al. (1998). Although initially focused on manual writing tasks, it established the necessary fundamentals for future research in the field. In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton presented AlexNet Krizhevsky et al. (2017). This model was submitted to the ImageNet competition Deng et al. (2009). ImageNet is a large image database. This dataset has been widely used in deep learning, containing millions of labeled images covering thousands of categories. It provides a valuable set for training and evaluating algorithms focused on object recognition. The ILSVRC (ImageNet Large Scale Visual Recognition Challenge) was created to compare different CV (Computer Vision) algorithms using this dataset. The AlexNet model won, beating the second-place winner by 10%. The architecture of this model is based on the proposed evolution of LeNet Lecun et al. (1998), with several significant improvements. Discarding more complex and abstract features was possible using stacked CNN (Convolutional Neural Networks) layers. MP (Max Pooling) layers played a critical role in feature selection, determining only the most relevant features, which would propagate deeper into the network. Images are classified into their respective categories using densely connected layers. A visual representation of the AlexNet structure is provided in Figure 2.12 for a better understanding.

Due to the good results of such models, combining multiple levels of CNN (Convolutional Neural Networks), along with MP (Max Pooling) layers and dense layers, has established the basis for convolutional network approaches in CV (Computer Vision). Existing foundations and models that solve more complex problems are examined in 2.3.2 and 2.3.3.

2.3.2 Object Detection

Object detection is a significant evolution from classifying objects in an image. Unlike classification, which is limited to labeling the appropriate category according to a single object in the image, object detection involves **identifying** and **precisely localizing multiple objects** through bounding boxes, represented as:

$$\text{bbox}_i = (x_i, y_i, \text{width}_i, \text{height}_i) \tag{2.24}$$

Where x_i, y_i are the minimum coordinates on the horizontal and vertical axes. width_i and height_i set the width and length of the bounding box. Figure 2.13 shows

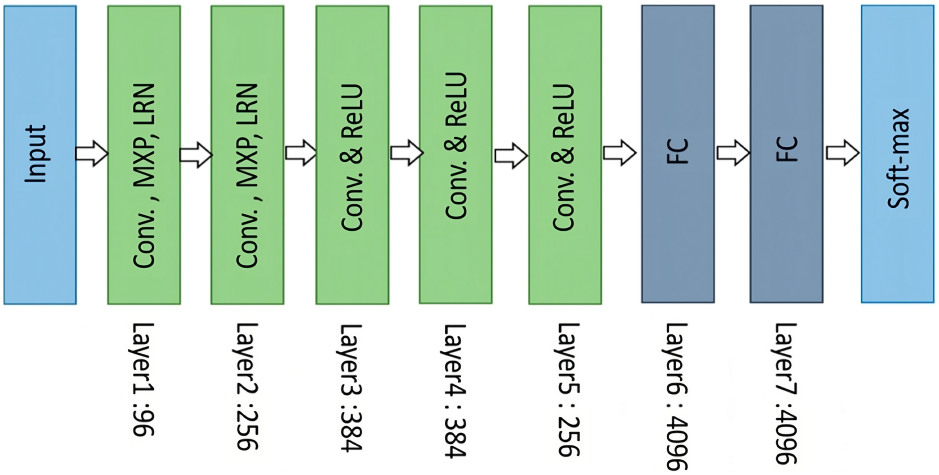

Figure 2.12: Architecture of the AlexNet model using an image from Alom et al. (2018).

an object detection example.

Figure 2.13: An object detection example using an image from the UAVDT Dataset Du et al. (2018).

This technique provides a more **detailed understanding** of the image composition, which benefits multiple areas, such as security in video surveillance systems. Several approaches have been developed over the years to address the task of object detection, resulting in two main categories: single-stage models and multi-stage models. These models are introduced in the subsections 2.3.2.1 and 2.3.2.2, respectively. The performance of the most prominent models, as well as their limitations, are also discussed.

2.3.2.1 One-Stage Object Detection

This type of model simplifies the detection task. The key is to exploit all the advantages of CNN (Convolutional Neural Networks) to detect the classes of elements and their locations in a **single step**. The main idea is based on using a model to classify the elements and then removing the last dense layers and some of the convolutional layers to obtain the result of a layer with a desired size. The activations of this layer are used to infer both the class and the element's location. Instead of grouping all the information into a single dense layer, as in the classification problem, convolutional layers are used to obtain the class, its BBOX (Bounding Box), and the probability that an object is found in that grid. Several models have emerged in this area. Each model has a specific performance, leading to some advantages and limitations. These are defined below:

- **CenterNet KPTS:** The CenterNet model focuses on feature recognition by predicting the center points of features and their respective bounding boxes. For each point in the feature map, a number C_i of confidence maps is generated, where each i represents a particular class. Each confidence map represents the probability that the center point belongs to a specific class. Then, the associated BBOX (Bounding Box) is predicted for each **detected point**. This is done by estimating the displacements from the central point to the corners of the BBOX (Bounding Box), represented as follows:

$$B_i = (x_i, y_i, w_i, h_i) \tag{2.25}$$

  Where B_i is the BBOX (Bounding Box) of the object i. (x_i, y_i) are the center coordinates, while (w_i, h_i) are the width and height predictions set for the bounding box.

  The main advantage of this model is easy object detection. There is no need to generate a region proposal, being able to detect objects of different shapes and sizes accurately. However, it can have problems identifying partially occluded or overlapping objects. Figure 2.14 shows the structure of the model.

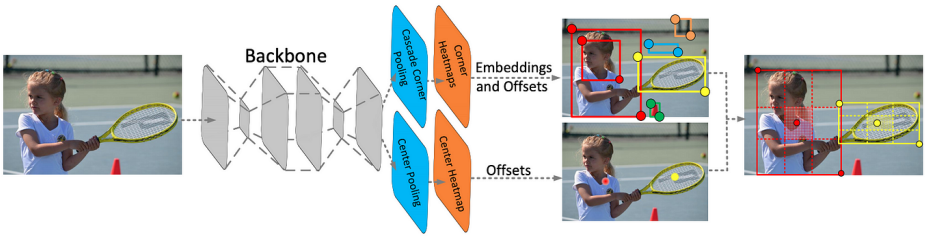

Figure 2.14: Structure of the CenterNet model using an image from Duan et al. (2019).

- **EfficientDet**: A family of models based on the EfficientNet architecture Tan and Le (2019). It uses a **compound scaling method** to achieve optimal efficiency. The recognition process consists of several steps Tan et al. (2020).

First, the network architecture is optimized through compound scaling, efficiently increasing the network depth, width, or resolution. Next, this family of models uses multiple **levels of resolution** in the feature pyramid to identify objects of different sizes, applying convolutions at different feature scales. For each resolution level, separate BBOX (Bounding Box) and class predictions are made using several additional convolution layers previously adapted to each scale. Finally, after obtaining the predictions, the NMS (Non Maximum Suppression) operation is applied, and the detections are filtered according to a specific confidence threshold.

The main advantage of this model family is that it provides an **optimal balance** between performance and computational efficiency. It also scales efficiently to identify objects of different sizes. As a disadvantage, this type of network can be less accurate than more complex models.

- **YOLO:** The YOLO (You Only Look Once) model is one of the most famous models of the one-step group. Given an input image I, YOLO (You Only Look Once) divides the image into a **grid of cells**, each representing a region of the image to be inferred. Each cell B predicts a set of bounding boxes and C object classes. These predictions include BBOX (Bounding Box) information such as coordinates, width, and length, as well as the confidence probability P_c that an object is present in that box, along with P_{class} for each of the C classes. The coordinates of the bounding boxes are transformed relative to the cell grid and the original image's size to obtain their exact location in the image. Finally, similar to other models, the NMS (Non Maximum Suppression) operation is applied with a previously defined threshold.

  The main advantage of this model is its speed and efficiency, which allows it to be used in areas where real-time response is required, providing accurate object detection in a single pass. However, depending on the cell grids' resolution, it can have difficulty detecting very small objects. It can also have problems detecting extremely large or overlapping elements. Today, this model has evolved into several versions, the most prominent of which is the V8 model Jocher et al. (2023). The architecture of this model is shown in Figure 2.15.

2.3.2.2 Multi-Stage Object Detection

The models in this group represent a classical and efficient approach. Unlike the one-stage models described in 2.3.2.1, the two-stage models divide the detection task into two distinct stages:

1. Regional proposal generation: A model is applied to determine **potential candidate regions** containing items.
2. Object recognition: An object recognition model is used for each candidate region identified above to determine which objects are found in each region.

This subdivision results in higher recognition accuracy. Within this group, there are several models, such as RCNN (Regions with CNN Features) Girshick et al. (2014), FASTRCNN (Fast Region-based Convolutional Network) Girshick (2015), or FASTERCNN (Faster Region-Convolutional Neural Network) Ren et al. (2017). The latter is described in detail below.

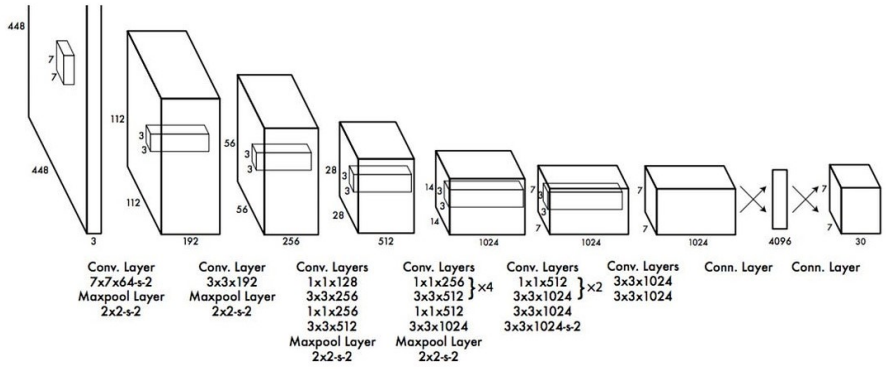

Figure 2.15: Architecture of the YOLO model using an image from Redmon et al. (2016).

- Faster R-CNN: In the first stage, a network known as the RPN (Region Proposal Network) generates a set of **candidate regions** in the input image where items may be contained. The RPN (Region Proposal Network) is responsible for examining the extracted features in the image. A convolution process determines a set of proposed bounding boxes representing the regions. In the second stage, each of the generated proposals is matched to the features of the original image using a layer known as RoI Pooling to predict the classes and regions of the bounding boxes of each proposal.

  The advantage of this type of model is its high accuracy in object detection, which allows objects of different sizes to be effectively identified. RPN (Region Proposal Network) eliminates the need to inspect the entire image, which improves efficiency. However, this approach introduces greater complexity than one-stage models, increasing the required training time and computational load. As a result, its implementation in real-time applications is limited because the model is slower regarding inference speed. Figure 2.16 visualizes the architecture of this model.

2.3.3 Object Segmentation

Object recognition identifies the object with a BBOX (Bounding Box). However, this information is **still inaccurate**. The Bbox frames the object but does not determine its shape. There are cases where the object does not have square or rectangular proportions. Therefore, the next step is to get this information **more accurately** at the **pixel level**.

Object segmentation has diversified into three different approaches that are explored below: pixel-level segmentation, semantic segmentation, and instance segmentation. Each approach addresses specific aspects of the field and provides solutions to specific challenges.

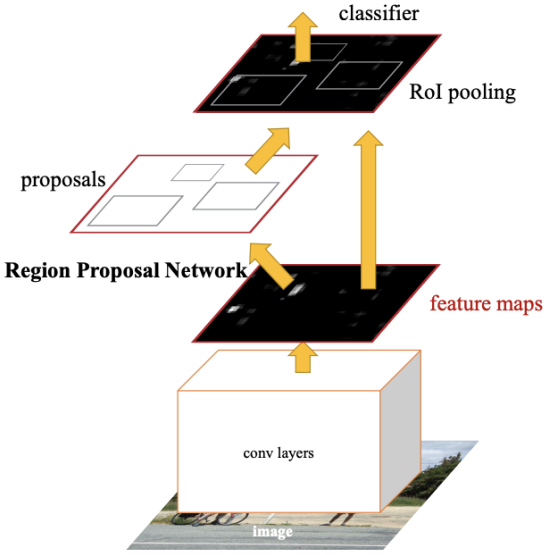

Figure 2.16: Architecture of the Faster R-CNN model using an image from Ahmed et al. (2021).

2.3.3.1 Pixel-Level Segmentation

The principle of pixel-level segmentation is based on assigning a class label to each pixel in an image. It provides a detailed understanding of the location of elements because each pixel is individually labeled. This results in a pixel segmentation mask, a visual representation that precisely determines the location of the elements in the image.

This technique is used in situations or areas requiring a high level of detail, such as identifying anomalies in magnetic resonance images in the clinical field. It differs from other segmentation methods, such as semantic or instance segmentation, in its **level of granularity**. However, effective training of this model type may require computational resources and large labeled datasets due to the high level of accuracy.

2.3.3.2 Semantic Segmentation

Semantic segmentation is based on assigning a **class label** to **each pixel** in an image, with the peculiarity of **not distinguishing each object**. This technique is based on understanding the categories of elements in an image as a group. It is designed to focus on the interpretation of visual scenes. Given a function S, it will be responsible for the mapping between the input image and the semantic segmentation map. Some of the applications of this type of technique are, for example, the implementation in the field of autonomous driving to understand the environment surrounding the vehicle and recognize the presence of various elements such as traffic signs, pedestrians, and other vehicles. An example of semantic segmentation is shown in Figure 2.17:

One of the best-known models in this area is U-Net Ronneberger et al. (2015),

Figure 2.17: Example of the Semantic Segmentation of a scene from the Cityscapes dataset Cordts et al. (2016).

a deep neural network with an **architecture** that combines feature encoding and decoding to obtain accurate semantic segmentation. While encoding progressively reduces the image, decoding reconstructs it using upsampling techniques to preserve fine details such as high-level features. Its adaptability has promoted its use in several domains, making it a reference model. Figure 2.18 visualizes the architecture of this model.

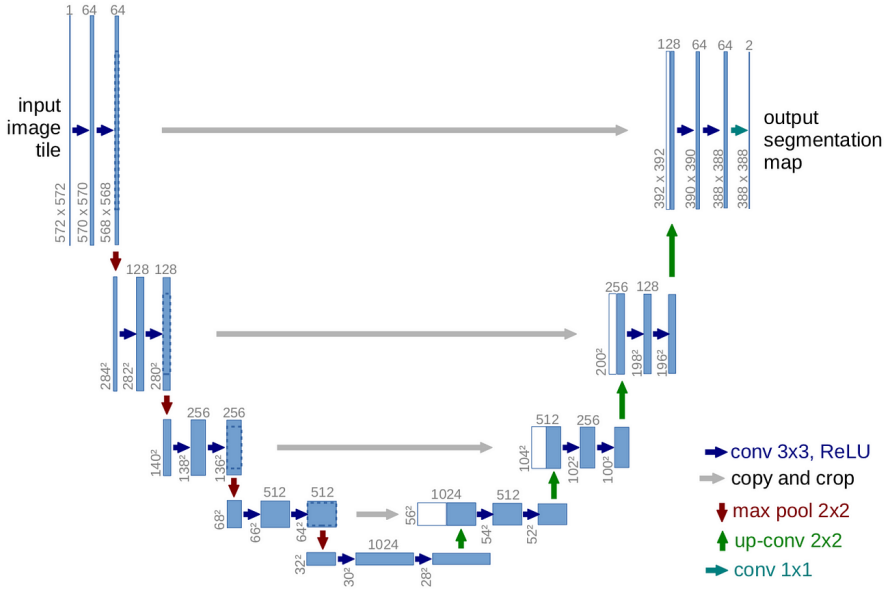

Figure 2.18: Architecture of the U-Net model using an image from Ronneberger et al. (2015).

This approach has several limitations and challenges since it is not possible to distinguish each object of the same type individually, leading to the development of more advanced techniques to solve the problem, such as instance segmentation.

2.3.3.3 Instance Segmentation

Instance segmentation is an advanced technique in CV. It is based on the combination of object detection and semantic segmentation capabilities. It is possible to **individually identify** and **separate each segmented object** and its respective class using this technique, assigning a unique identifier to each object instance. This technique is used in situations where high precision in the individual identification of each element is required. An example of instance segmentation is shown in Figure 2.19.

Figure 2.19: Example of instance segmentation using Detectron2 Wu et al. (2019).

Two widely recognized and effective models stand out in this area:

- MASK R-CNN: MASKRCNN (Mask Region-based Convolutional Neural Network) He et al. (2017b) is based on extending the FASTERCNN (Faster Region-Convolutional Neural Network) architecture to include mask generation. First, MASKRCNN uses an object detection network to determine a set of bounding boxes and class probabilities. Then, a mask network is employed for each detected object to generate accurate segmentation masks. The output of this model is the combination of the object detections and their corresponding segmented masks. Figure 2.20 shows the model's architecture.
- YOLACT++: YOLACT++ (You Only Look At CoefficientTs ++) Bolya et al. (2022) is an extension and improvement of the original YOLACT model Bolya et al. (2019). Its operation is based on decomposing spatial and channel features, allowing a precise segmentation of objects. It uses a set of anchors called A to predict each object's Bbox coordinates B and the class probability C individually. Then, instead of using the direct attention mask as its predecessor, it uses coefficient features, denoted K. These are multiplied by the extracted features to generate accurate segmentation masks. Finally, the

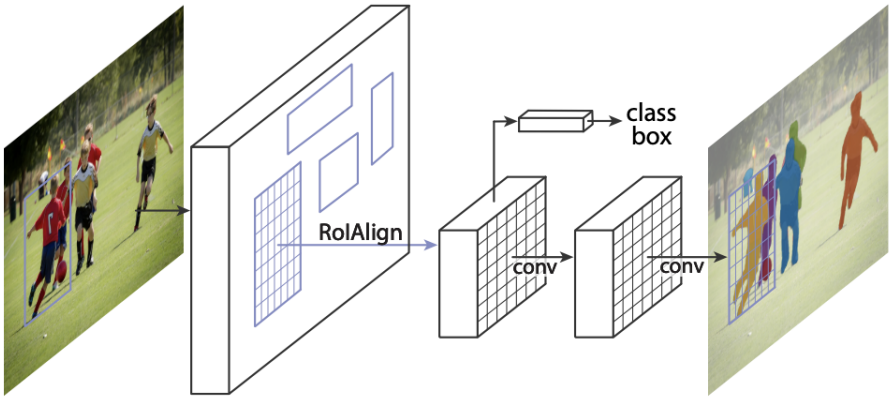

Figure 2.20: Architecture of the MASK R-CNN model using an image from He et al. (2017a).

output combines the object detection and the coefficient-based segmentation masks. The architecture of such a model is shown in Figure 2.21.

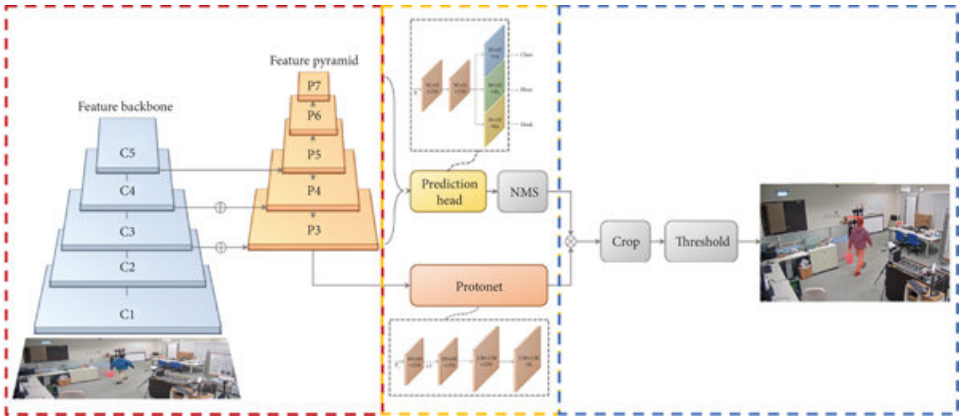

Figure 2.21: Architecture of the YOLACT++ model using an image from Tseng et al. (2021).

Although instance segmentation aims to achieve a high level of detail when identifying elements, it has multiple limitations. One associated problem is based on several elements overlapping at the same location in the image. Another problem is based on the detection of small or partially visible objects.

2.3.4 Introduction of small objects

The object detection problem has experienced a notorious advance promoted by DL. It is now possible to autonomously analyze and understand images and video sequences with good accuracy. However, there are still several unsolved challenges

that need to be improved. One is the efficient detection and segmentation of **small objects**. Due to their reduced size, small objects present several problems, such as the lack of distinctive features, which are challenging to identify.

This thesis presents a series of works that aim to mitigate this problem as much as possible, thereby increasing the mAP of models based on CNN. For this purpose, this subsection examines the problem and possible solutions that can be implemented to mitigate it. It also discusses the main drawbacks and difficulties, the applications that can benefit from this development, and some existing proposals.

2.3.4.1 The Problem

The problem lies mainly in the difficulty of detecting and segmenting small objects in an input image. This is a **significant challenge** in CV because their characteristics are usually difficult to distinguish from the surrounding elements. Generally, a small object is considered an entity or element with significantly reduced dimensions compared to the average size of other objects in the same scene. The characteristics of such elements can be diverse in nature and shape. However, their size is much smaller concerning the surrounding environment. Depending on the application and the scale of the image, this size can differ.

The main objective is to provide methods to improve the detection and segmentation of mainly small objects. Currently, many models provide good results when identifying elements of common size. However, the **accuracy decreases** when the **size** of these elements is **reduced**. Most of these models reduce the resolution of the images in the intermediate layers that conform to the image, causing the loss of features of these elements, which end up disappearing in the processing of the network. All this, together with the problems presented previously, makes detecting this type of element a **complex task**.

An example would be detecting cars or pedestrians in urban sequences by video surveillance systems. Due to these characteristics, it can be difficult to identify and segment them accurately. Figure 2.22 shows an example of the shortcomings of small object detection using CNN.

2.3.5 Solution

To address the challenge of small object detection and segmentation, it is important to determine the problem's complexity and the possible solutions to mitigate it. It is useful to consider several **fundamental questions** to evaluate any proposed method:

- How is the initial detection of objects performed?: Determine the method responsible for their identification, for example, the use of pre-trained models. It is also necessary to determine how variability in size and scale is handled.
- How is object segmentation performed?: It is necessary to determine the segmentation approach that will be performed, such as contour-based and other specific strategies. In addition, it is important to define how to handle partial occlusion challenges to ensure accuracy and robustness.
- Are any pre-processing methods used?: Specify if any techniques have been used to improve the quality of the input image.

Figure 2.22: Example of the shortcomings of small object detection using convolutional neural networks selecting an image from the NGSIM (Next Generation Simulation) Dataset U.S. Department Of Transportation Federal Highway Administration (2017).

- How are the challenges posed by identifying this type of elements addressed?: Define how the robustness of the method is ensured in a variety of specific scenarios and conditions.

These questions provide a starting point for evaluating the effectiveness and suitability of small element identification methods. Many approaches have been established to address this problem as much as possible:

- Hyperparameter tuning: Based on the optimization of hyperparameters of the selected model, including variables such as learning rate, activation functions, network architecture, and others.
- Postprocessing Techniques: Apply techniques to refine and improve the identification of these elements by the CNN, such as contour fitting.
- Apply Fine-Tuning or Transfer Learning: Starting from pre-trained CNN models, refine them using a dataset specialized in detecting and segmenting small objects to be detected, thus adapting them to this task.
- Data Augmentation: Its application in combination with CNN generates a series of variants of the original image with a series of changes such as scale or

orientation, allowing the networks to learn to detect objects under different conditions.

- Use of specialized architectures: Some approaches are based on customized networks for a particular problem. For this purpose, architectures are designed with layers and blocks specifically designed for this purpose.
- Incorporation of other mechanisms: Attention and contextual approaches allow the model to focus on a specific image area where the relevant small elements to be identified are usually located.

According to the usual approaches listed above, they help deal with the complexity of the presented problem based on the identification of elements with reduced size. However, despite these techniques, there is still a need for new solutions to mitigate the problem.

2.3.6 Commom Problems to Face

Small objects, characterized by their small size, present unique characteristics and problems that prevent them from properly identifying. The specific problems are listed below:

- Varying sizes and scales: Within a given application, objects can have a variety of sizes and scales. An object can be very small concerning the image in which it is located, making it susceptible to being overlooked.
- Deficiency of features: Based on the lack of features that differentiate the object from the background or other elements. These distinguishing features include a lack of sharp details, distinctive textures, or sharp edges.
- Complex environments: Influences whether the objects to be identified are in complex and noisy environments, leading to false detections or even class confusion. Background elements such as specific colors or patterns can affect the robustness of the applied algorithms.
- Object occlusion: Both overlapping and occlusion can interfere with identifying such elements, which may be partially hidden or overlapping.
- Image quality: The resolution and sharpness of the input data are critical features. Another aspect to consider is the distance at which the elements were captured. Robustness will decrease the lower the level of detail of the scenes to be analyzed.
- Adverse conditions: Lighting, backlighting, reflections, or shadows can affect the visibility and appearance of objects.
- Similarity: The presence of nearby elements with similar shapes and textures can lead to confusion in correctly identifying them.

These problems represent the most important challenges in improving small element identification. An example of the problems listed above is shown in Figure 2.23. The challenges presented above require new techniques or innovative solutions to mitigate them to facilitate their use in various areas.

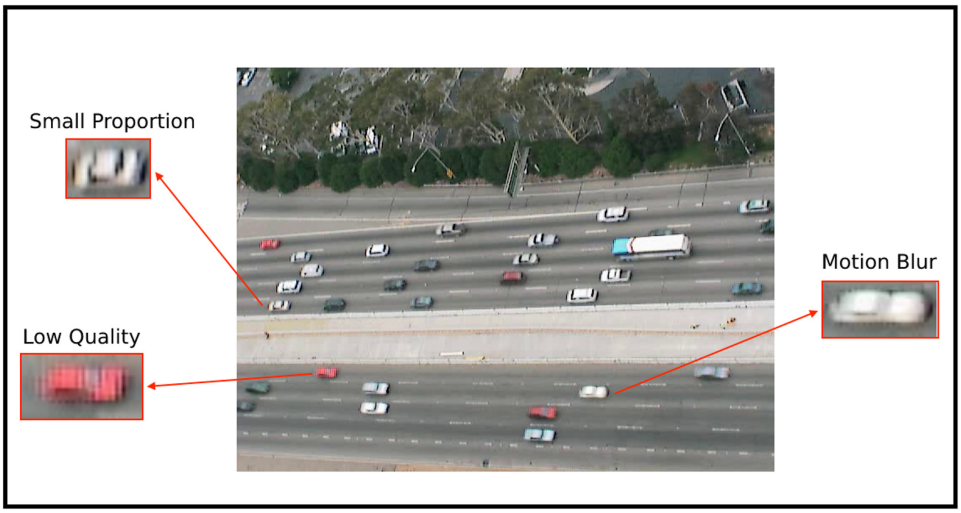

Figure 2.23: Example of the problems to detect small objects selecting an image from the NGSIM (Next Generation Simulation) Dataset U.S. Department Of Transportation Federal Highway Administration (2017).

2.3.7 Applications

Several areas require solutions to mitigate the identification of small elements that benefit in multiple ways. Some of these are:

- Surveillance and security: Applying this technology to video surveillance systems can be critical for identifying objects, such as detecting intruders in sensitive areas.
- Autonomous Vehicles: In the automotive industry, to ensure the safety of autonomous vehicles and assisted driving systems, the detection of elements such as other vehicles or pedestrians is essential. Identifying them helps avoid accidents and ensures safe driving in many different environments.
- Medical and medical diagnostics: Detecting or segmenting small objects is fundamental in medicine. For example, such techniques could be used to detect cancer cells in histology for early and accurate diagnosis of many diseases. For specific tasks, identifying objects in images is crucial for adequately controlling disease and for medical follow-up.
- Industrial quality control: Detecting small defects in electrical, textile, or pharmaceutical components contributes directly to the quality of the product and the company's production efficiency.
- Robotics: In applications focused on automation and robotics, the ability to detect small elements allows tasks such as the handling of small objects, where processes, where precision is critical, can benefit.
- Agriculture: The correct identification of this type of element can support and help in the early detection of diseases or pests, thus helping to take preventive measures and improve agricultural management.

These applications cover a wide range of fields, thus opening up new opportunities for research and technological development in various sectors. In this thesis, we present a series of works that support them. These works are mainly focused on video surveillance sequences.

2.3.8 Some Previous Proposal

Small object detection is a relatively recent problem. In its early stages, the task was approached through the features of the surface elements and the classifiers themselves. Vehicle detection in aerial imagery was the main focus of these tasks. Kembhavi et al. (2011) determined a methodology where, from heat probability maps, they capture statistics of vehicles and their environment, oriented gradient histograms, and an image descriptor that takes the name of pixel pairs responsible for capturing the structural features of objects. Other advances focus mainly on the color and shape of the elements to be detected, such as Le et al. (2010) for traffic sign detection.

The advancement of CNN has led to several methods, mainly focused on DL, which have experienced significant improvements using them. Regarding object detection methods based on DCNN (Deep Convolutional Neural Networks), intense research has emerged in recent years, with pioneering works such as Girshick et al. (2014); Girshick (2015), where simple and scalable detection algorithms that improve accuracy are proposed based on the joint application of regional proposals together with deep networks, leading to increasingly sophisticated models and proposals Ren et al. (2017); Liu et al. (2016); Tan and Le (2019); Zhu et al. (2021). The close relationship between detection and object segmentation solves both by addressing similar weaknesses and problems. Thus, several models and strategies have been developed over the years for object segmentation as a natural evolution of object detection. Examples include the MASKRCNN (Mask Region-based Convolutional Neural Network) model He et al. (2017b), which extends the FASTER-CNN (Faster Region-Convolutional Neural Network) detection model Ren et al. (2017) by adding a parallel branch to the bounding box extraction for the mask prediction of each object. Proposals such as Hayder et al. (2017), base their segmentations on the shape of objects to reduce the dependence on the quality of the Bbox and thus avoid failures caused by erroneous proposals inherited from object detection.

However, in the small object detection field, relatively few studies focus on small object detection. Chen et al. (2017) address this problem by establishing a focused SOD (Small Object Detection) dataset and evaluation metrics to determine a benchmark in the field. Krishna and Jawahar (2017), based on the ideas established by Chen et al., propose a technique based on bottom-up sampling that provides better results. Over the years, the methods developed for this problem have mainly been based on creating specific architectures and models for a particular model or applying techniques that support improving object detection and segmentation algorithms. Regarding the second category, multiscale methods stand out, such as the one proposed by Lin et al. (2017), where they present the FPN (Feature Pyramid Network), which uses deep convolutional networks to build these pyramids efficiently, thus improving object detection.

There are also other approaches based on DA (Data Augmentation) techniques, such as the one proposed by Kisantal et al. (2019), where images are oversampled

with small objects and augmented by copying and pasting these objects multiple times to improve the quality of the detection model. Akyon et al. (2022) propose a framework known as SAHI (Slicing Aided Hyper Inference). This generic solution is based on segmentation-assisted inference, keeping complexity and memory requirements low to improve performance, using data augmentation techniques and retraining strategies. Within the field of strategies based on image segmentation, approaches such as Laradji et al. (2019) perform training from image-level labeled data by obtaining pseudo-masks before training a MASKRCNN (Mask Region-based Convolutional Neural Network), or Laradji et al. (2020), which requires the annotation of a single pixel for each object in the image. Other proposals are based on applying techniques that improve the performance of deep networks. Yang et al. (2020) improves the MASKRCNN (Mask Region-based Convolutional Neural Network) model using a multi-scale region proposal network structure. Zhang et al. (2020) conducts a study of the relationship between the spatial information of the receptive field and the size of the object with the segmentation and proposes MRRCNN (Multi-Region CNN), a network with convolutional layers attached to the proposed semantic segmentation layer containing feature pyramids. One of the lines is supported by the use of artificial data generation. Ghiasi et al. (2021) generate synthetic data based on cutting and pasting instances of objects in other images.

Recently, the use of GAN (Generative Adversarial Networks) has made a significant advance with works such as the one proposed by Bai et al. (2018), where an approach called MTGAN (Generative Adversarial Network) is presented. The generator is responsible for improving the quality of the images, and the discriminator performs classification and detection tasks.

2.4 Fundamentals of Anomaly Detection

Urban mobility is an essential aspect that affects many areas, including the economy and the quality of life of its inhabitants. Several complex challenges need to be addressed in the field of pedestrian and vehicle flow management. One of them is the **detection of anomalies** in the flow of cars and pedestrians, which significantly impacts the safety, planning, and efficiency of cities. Some works focused on detecting anomalies within this field are provided to support this thesis. Therefore, the basics are presented in the following points, identifying the problems and possible solutions to deal with them. Some recent advances attempting to mitigate this problem are also described.

2.4.1 What is an anomaly

An anomaly in urban traffic is an **unusual behavior or deviation** in vehicle or pedestrian traffic significantly different from the typical traffic norms and patterns established in a given area. Mathematically, this can be defined as:

$$|x_i - \mu| > k \cdot \sigma \tag{2.26}$$

Where $|x_i - \mu|$ is the difference between the observation distance x_i and the mean μ of the data. K is used to determine how many standard deviations from

a mean are considered anomalous, where *sigma* is the standard deviation of the data.

There are several anomalies within this range. Some examples are

- Circulation in forbidden areas: This situation occurs when a car circulates or parks in prohibited areas, such as a crosswalk. In this case, it is considered an anomaly because it obstructs traffic and does not correspond to an area where vehicles normally park.
- Crosswalk in prohibited areas: As far as pedestrians are concerned, the passage of pedestrians outside a dedicated pedestrian crossing is considered an anomaly because it increases the risk of accidents.

These examples illustrate some of the possible anomalies that can occur in traffic areas, increasing the risk of aspects related to road safety or traffic flow.

2.4.2 Problems and Solutions

Several video surveillance systems at high and strategic points manage security in urban environments. These cameras have poor quality in certain situations. In addition, they face many challenges that make it difficult to identify certain anomalies. The manual identification of such anomalies is **very expensive**, as it would require manual identification by several workers. The number of operators needed for this purpose would be exponential concerning the number of cameras under analysis, making this task infeasible. For this reason, it is necessary to develop solutions that allow the same task to be performed autonomously. However, it is required first to face some challenges that correspond to the following:

- Difficulty in identifying small objects: The problem of identifying small elements such as pedestrians or vehicles arises again because the sequences are captured by cameras located at high points. As a result, some elements are composed of a low pixel rate.
- Loss of visual detail: The distance between these video surveillance systems and the captured elements can lead to a loss of detail. Weather also comes into play, with low visibility situations such as fog, rain, or lightning in high-brightness scenes.
- Expensive data annotation: Obtaining labeled data sets in which anomalies of various sizes are manually identified is costly and time-consuming, severely limiting the capacity required for model training.

Given the difficulties described above, several possible solutions have been proposed to mitigate this problem. Some of these are:

- Detection of objects in areas of interest: The anomaly detection model is then applied to these areas to reduce the computational load and focus on the most vulnerable areas where anomalies may occur by using detection systems that correctly identify points of interest, such as roads or intersections.
- Generation of synthetic datasets: Using simulators such as CARLA (Car Learning to Act) Dosovitskiy et al. (2017), a series of synthetic images can

be generated to expand the dataset. Their annotations can be obtained autonomously. The model would then be trained to detect anomalies.

- Supervised Learning: Semi-supervised learning techniques use unlabeled data for anomaly detection to help the model detect specific patterns not present in the labeled data.

These solutions can mitigate specific problems in anomaly detection using CNN. Developing techniques related to these approaches would lead to methodologies that significantly improve the capability of surveillance systems in urban environments.

2.4.3 Some Previous Proposals

The first early techniques that addressed anomaly detection using classical techniques relying on reconstructive or discriminative approaches Sabokrou et al. (2015); Cong et al. (2011); Zhao et al. (2011); Mahadevan et al. (2010). The objective was to learn the standard behavior patterns in a specific domain, getting good results. However, there are several limitations in capturing complex distributions in video sequences. CNN application to anomaly detection has resulted in research like the one introduced by Hu et al. (2020). Their work proposes a weakly supervised framework for identifying and localizing unusual behavior in scenes. This framework employs object detection using FASTERCNN (Faster Region-Convolutional Neural Network), behavior description through a LSOF (Large Scale Optical Flow) Histogram descriptor, and classification with a MISVM (Multiple Instance Support Vector Machine). Kanu Asiegbu's research in 2021 Kanu-Asiegbu et al. (2021) presents an innovative approach to the unsupervised detection of anomalous pedestrian events. Unlike traditional reconstruction-based methods, their framework leverages prediction errors from normal and abnormal pedestrian trajectories to detect spatial and temporal anomalies. Empirical results on real-world benchmark datasets showcase the effectiveness and efficiency of this trajectory-based anomaly detection method in identifying unusual pedestrian activities across different timeframes.

In studies like the one referenced as Chang et al. (2022), a proposed deep learning model for abnormal behavior detection employs YOLO (You Only Look Once) v3 Redmon and Farhadi (2018) object detection technology to identify pedestrians. Subsequently, it utilizes a hybrid DEEPSORT (Deep Simple Online Realtime Tracking) algorithm to track pedestrians and capture their trajectories. A CNN also extracts action features from each tracked trajectory. In contrast, a LSTM (Long short-term memory) is employed to establish a model for identifying and predicting anomalous behaviors. Another approach, described in Gayal and Patil (2023), presents an automatic anomaly detection model that combines hierarchical social hunting optimization with a deep convolutional neural network (HiS-Deep CNN) for analyzing surveillance videos. This model encompasses object detection and tracking. Nevertheless, some of these methodologies face the challenge of acquiring sufficient training data. Therefore, the methodology presented in this article utilizes synthetic data generated through the CARLA (Car Learning to Act) simulator Dosovitskiy et al. (2017). Furthermore, B. Sophia (2023) introduce a novel PFPN-ADT (Panoptic Feature Pyramid Network-based Anomaly Detection and Tracking) model designed for pedestrian walkways in video surveillance. The primary objective of this model is to identify and classify anomalies in pedestrian walkways, such

as vehicles and skaters. Their approach leverages the PFPN (Panoptic Feature Pyramid Network) for object recognition. It combines it with the CBA (Compact Bat Algorithm) and a SAE (Stacked Auto Encoder) for object classification. This demonstrates improved performance in effectively detecting anomalies using the PFPN-ADT (Panoptic Feature Pyramid Network-based Anomaly Detection and Tracking) technique.

Due to the growing demand for enhanced security and safety measures, there has been a notable interest in intelligent video surveillance analysis. In Shen et al. (2017), a pioneering method for detecting anomalies in pedestrian behavior is proposed. This approach leverages motion-appearance features and dynamic behavioral changes over time, utilizing LSH (Locality Sensitive Hashing) functions for detection. Key contributions include robust pedestrian segmentation, the DoPB (Dynamics of Pedestrian Behavior) feature, and the AAW (Adaptive Anomaly Weight) combined with block-based optical flow tracking, demonstrating its efficacy in detecting and pinpointing anomalies.

Additionally, Irina et al. introduce an automated anomaly detection technique based on deep learning, known as DLADT-PW (Deep Learning based Anomaly Detection Technique in Pedestrian Walkways) Pustokhina et al. (2021), to enhance pedestrian safety. The conventional manual inspection of abnormal events within video surveillance systems is labor-intensive, underscoring the importance of automated surveillance systems for computer vision researchers. DLADT-PW (Deep Learning based Anomaly Detection Technique in Pedestrian Walkways) incorporates preprocessing to eliminate noise and enhance image quality. The detection process involves the utilization of the MASKRCNN (Mask Region-based Convolutional Neural Network) He et al. (2017a) in conjunction with DenseNet (Densely Connected Networks). The DLADT-PW (Deep Learning based Anomaly Detection Technique in Pedestrian Walkways) model is designed to detect and classify anomalies in pedestrian walkways, such as vehicles, skaters, and jeeps.

Chapter 3

Improved detection of small objects in road network sequences using CNN and super resolution

It always seems impossible until it's done.
Nelson Mandela.

Abstract: This first chapter of the research work presented in this Ph.D. thesis includes a paper that was published in the journal EX (Expert Systems) in the year 2021. The research addresses the significant challenge of detecting small objects in DL, where contextual complexity or limited pixel count often leads to suboptimal performance. A proposed novel methodology uses pre-trained CNN models to detect vehicles in images captured by video surveillance cameras without modifying the network structures or retraining them to address this issue. The approach goes beyond traditional methods by improving object detection and class inference from specific initial regions while seamlessly integrating image resolution enhancement processes. Multiple tests on a diverse set of traffic images with objects of different scales validate the effectiveness of our solution. Notably, the proposal consistently outperforms the pre-trained models, achieving an impressive average accuracy of 45.1% with the EfficientDet-D4 model for the initial video sequence, a significant improvement over the model's original accuracy of 24.3%.

Title	Improved detection of small objects in road network sequences using CNN and super resolution
Authors	Iván García-Aguilar, Rafael M. Luque-Baena, Ezequiel López-Rubio
Journal	Expert Systems
Year	2021
Impact Factor	2.812
JCR Categories	COMPUTER SCIENCE, THEORY & METHODS (37/110) (Q2)
Status	Published
DOI	https://doi.org/10.1111/exsy.12930
Cite	García-Aguilar et al. 2022

Chapter 4

Enhanced Image Segmentation by a Novel Test Time Augmentation and Super-Resolution

Progress is impossible without change,
and those who cannot change their
minds cannot change anything.

George Bernard Shaw.

Abstract: This chapter presents our work to address the challenge of image segmentation in computer vision applications. Presented at the 9th IWINAC (International Work-Conference on the Interplay Between Natural and Artificial Computation) in Tenerife, Spain, in 2022, this paper outlines a novel methodology to improve the accuracy and efficiency of segmentation. In real-world scenarios where dynamic elements interact with moving backgrounds, achieving accurate results requires sophisticated models. Our approach uses SR techniques and re-inference to improve image segmentation quality without requiring model modifications or retraining. We demonstrate the significant improvement achieved by our approach and establish its potential as a valuable tool in computer vision through a series of experiments.

Title	Enhanced Image Segmentation by a Novel Test Time Augmentation and Super-Resolution
Authors	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio, Enrique Domínguez
Conference	9th INTERNATIONAL WORK-CONFERENCE ON THE INTERPLAY BETWEEN NATURAL AND ARTIFICIAL COMPUTATION (IWINAC)
Year	2022
GGS Rating	Work in Progress
CORE Rating	-
Status	Published
DOI	https://doi.org/10.1007/978-3-031-06527-9_15
Cite	García-Aguilar et al. 2022

Chapter 5

Optimized instance segmentation by super-resolution and maximal clique generation

I could never resist the temptation of having a look at something that doesn't exist.

The Last Wish, Andrzej Sapkowski.

Abstract: This chapter introduces a paper presented to the journal ICAE (Integrated Computer-Aided Engineering) in 2023. It determines an innovative approach to enhance the efficiency of instant segmentation models in the context of an on-board camera. With the rapid proliferation of surveillance technology, there is a growing need to automate tasks within Intelligent Vehicle Systems and Intelligent Transport Systems, particularly in vehicle detection for traffic management, accident detection, and risk mitigation. This work proposed a meta-method that offers a versatile solution applicable to different instant segmentation models, such as MASKRCNN (Mask Region-based Convolutional Neural Network) and YOLACT++ (You Only Look At CoefficientTs ++). We achieve optimized re-inference by leveraging the initial detections generated by these models and integrating super-resolution techniques. This process enables the identification of objects that were initially missed and enhances the quality of detected elements. Our meta-method elegantly optimizes the required time to infer over the image by applying optimization techniques based on nearest-neighbor calculations. By strategically selecting regions to super-resolve, we minimize the number of images requiring re-inference. Experimental results validate the effectiveness of our proposal, showcasing improvements of up to 8.1% when applied to the

YOLACT++ (You Only Look At CoefficientTs ++) model in one of the sequences, significantly enhancing instant segmentation models in surveillance and intelligent transportation systems.

Title	Optimized instance segmentation by super-resolution and maximal clique generation
Authors	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio, Enrique Domínguez
Journal	Integrated Computer-Aided Engineering (ICAE)
Year	2023
Impact Factor	6.5
JCR Categories	COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (22/110) (Q1))
Status	Published
DOI	https://doi.org/10.3233/ica-230700
Cite	García-Aguilar et al. 2023b

Chapter 6

Automatic labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks

Wisdom is to have dreams big enough
not to lose sight when we pursue them.
Oscar Wilde.

Abstract:

This chapter presents a paper published in PRL (Pattern Recognition Letters) in 2023. Existing pre-trained models have certain shortcomings, especially in detecting small objects, leading to a lower recognition rate. A significant hurdle is the manual labeling required for each vehicle captured by many IP cameras distributed over extensive road networks in different regions. Our innovative proposal introduces an automatic procedure to detect small-scale objects within traffic sequences. In its initial phase, this procedure automatically detects vehicle patterns from a series of frames through an offline process. This process uses SR techniques in conjunction with pre-trained object detection networks. Subsequently, the object detection model is retrained using the previously obtained data, adapting to the unique characteristics of the analyzed scene. This dynamic adaptation aims to enhance the performance. The final stage of our framework takes place in real-time once the model has been pre-trained. This approach has been rigorously tested and validated, demonstrating its effective-

ness on popular datasets such as NGSIM (Next Generation Simulation) and GRAM. Our methodology represents a promising advance in addressing the challenges associated with real-time, small-scale vehicle detection and object identification in road management systems.

Title	Automatic labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks
Authors	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio
Journal	Pattern Recognition Letters (PRL)
Year	2023
Impact Factor	5.1
JCR Categories	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (52/145) (Q2))
Status	Published
DOI	https://doi.org/10.1016/j.patrec.2023.01.015
Cite	García-Aguilar et al. 2023

Chapter 7

Minimal Optimal Region Generation for Enhanced Object Detection in Aerial Images using Super-Resolution and Convolutional Neural Networks

Lack of something to feel important about is almost the greatest tragedy a man may have.

Red Dead Redemption 2, Arthur E. Morgan.

Abstract: The sixth research work chapter in this Ph.D. Thesis was presented at the 17th IWANN (International Work Conference on Artificial Neural Networks) in 2023, held in the Azores Islands, Portugal. This research paper presents a novel methodology to enhance object detection within aerial imagery. This novel approach combines two powerful techniques: SR and CNN. The workflow begins with segmenting the image's grey zone, a crucial step in identifying roads and other regions of interest. The YOLO (You Only Look Once) model is employed for this purpose, facilitating precise region identification. A new minimal optimal region was designed to encapsulate object instances. Within each of these regions, we leverage the power of SR techniques to enhance image clarity and resolution, effectively increas-

ing the number of pixels within the area to facilitate the application of CNN to these high-resolution regions. This approach improves object detection accuracy significantly, ensuring more precise localization and identification of objects within the aerial imagery, obtaining a substantial increase in mean average precision within the challenging domain of aerial imagery.

Title	Minimal Optimal Region Generation for Enhanced Object Detection in Aerial Images using Super-Resolution and Convolutional Neural Networks
Authors	Iván García-Aguilar, Lipika Deka, Rafael M. Luque-Baena, Enrique Domínguez, Ezequiel López-Rubio
Conference	17th International Work Conference on Artificial Neural Networks (IWANN)
Year	2023
GGS Rating	Work in Progress
CORE Rating	-
Status	Published
DOI	https://doi.org/10.1007/978-3-031-43085-5_22
Cite	García-Aguilar et al. 2023

Chapter 8

Object detection in traffic videos: an optimized approach using super-resolution and maximal clique algorithm

We can't solve problems by using the same kind of thinking we used when we created them.

Albert Einstein.

Abstract: This work was published in the journal *Neural Computing and Applications* in 2023 and represents a natural continuation of the research previously presented in chapter 7. We introduce an optimized approach based on deep object detection models, aiming to increase the number of detected elements and enhance the accuracy of their class inference. As a novelty, we compute a graph that describes the distances among the preliminary object detections and identify maximal cliques within it. Doing this reduces the number of windows requiring examination, significantly accelerating the detection process. Our framework has undergone rigorous testing on real traffic sequences from the U.S. Department of Transportation. The results are compelling, with an increase of up to 44.6% in detection rates, transitioning from an average detection rate of 14.5% for the EfficientDet D4 model to an impressive 59.1% when utilizing the methodology presented herein for the first sequence. In addition, qualitative experiments were conducted over the Cityscapes and VisDrone datasets, further validating the effectiveness of our approach and offering a promising solution to the persistent challenges associated with small object detection in deep

learning-based CV.

Title	Object detection in traffic videos: an optimized approach using super-resolution and maximal clique algorithm
Authors	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio
Journal	Neural Computing and Applications
Year	2023
Impact Factor	6.0
JCR Categories	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (41/145) (Q2))
Status	Published
DOI	https://doi.org/10.1007/s00521-023-08741-4
Cite	García-Aguilar et al. 2023a

Chapter 9

Small-Scale Urban Object Anomaly Detection Using Convolutional Neural Networks with Probability Estimation

Our imagination is the only limit to what we can hope to have in the future.

Charles Kettering.

Abstract: This chapter presents a paper published in *SENSORS* in 2023 that addresses anomaly detection within sequences, a critical challenge in security and surveillance. As surveillance cameras proliferate in urban road networks, the need for automated, efficient data analysis to detect anomalous events becomes essential. The methodology presented here offers a novel approach to detecting anomalies within urban sequences by harnessing the power of pre-trained CNN and SR models. The research unfolds in two distinct phases. In the offline phase, a pre-trained CNN model is used to evaluate a large dataset of urban sequences to identify and establish common locations associated with elements of interest. A density matrix that captures spatial patterns and highlights the most frequent element locations is also computed in this phase. Using the probabilities derived from the offline analysis, the pre-trained CNN moves to an online stage where it continuously evaluates real-time sequences using the density matrix to assess the likelihood of anomalies. As demonstrated in our experimental results, this dynamic approach enables the detection of various anomalies, including irregular pedestrian routes, providing a practical and reliable method to enhance public safety in urban environments.

Title	Small-Scale Urban Object Anomaly Detection Using Convolutional Neural Networks with Probability Estimation
Authors	Iván García-Aguilar, Rafael M. Luque-Baena, Enrique Domínguez, Ezequiel López-Rubio
Journal	SENSORS
Year	2023
Impact Factor	3.9
JCR Categories	ENGINEERING, ELECTRICAL & ELECTRONIC (100/275) (Q2))
Status	Published
DOI	https://doi.org/10.3390/s23167185
Cite	García-Aguilar et al. 2023

Chapter 10

Conclusions and Future Research Lines

Every day, we change the world, but to change the world in a way that means anything that takes more time than most people have. It never happens all at once. It's slow. It's methodical. It's exhausting.

Mr Robot, Elliot Alderson.

Abstract:

This chapter contains the conclusions from the seven works supporting this thesis and some future research lines motivated by the experimental results obtained after testing the different methods developed on them. The goal is to summarize this research period's main contributions and formulate some ideas for their further development toward more powerful and robust methods.

10.1 Conclusions

The main objective of this thesis has been the identification problem of small objects. It has been applied to both the detection and segmentation domains. This is a relatively new problem in the CV field. Therefore, searching for solutions to **mitigate** this problem is still in progress. Our approach is based on combining the use of techniques, and models focused on improving the image resolution to re-infer them according to different heuristics established to serve as a support to the pre-trained model of detecting or segmenting objects independently to improve the accuracy obtained, making most of the contributions centered on this approach. During this process, as knowledge was gradually acquired on different problems and areas, concepts emerged to use some of these methods in other areas, such as anomaly detection, representing the secondary part of this thesis.

In support of this research, seven papers have been included in which the author of the thesis is listed as the first author, except for the third. Five of these seven papers have been published in high-impact journals, and two smaller papers exploring preliminary ideas have been disseminated at lower-impact conferences.

The works that support this thesis are closely related, leading to the incremental development of a methodology to improve and apply it to different problems. Therefore, the papers supporting this thesis can be divided into three main groups. The first paper, published in Expert Systems (chapter 3), is the initial approximation of the described methodology to address the problem of small object detection. This work focuses on improving the accuracy of several pre-trained object detection models, applying a specific methodology based on re-referencing each super-resolved region generated from the initially detected elements in sequences where there are deficiencies in identifying all the elements that appear in the scene. The proposal consistently surpasses the performance of the pre-trained models, attaining a remarkable mean average precision of 45.1% using, for example, the EfficientDet-D4 model for the first video sequence. This represents a noteworthy enhancement compared to the model's original accuracy of 24.3%. This approach is used again in the work published in PRL (Pattern Recognition Letters) (chapter 6). In the first case, applying the methodology described made it possible to significantly improve the accuracy without modifying the model. However, the processing time of the scenes increased considerably since it was re-inferred to each of the initial elements recognized by the network. Therefore, in the second work, an additional step is applied in which the object detection model, taking into account the new elements identified with this methodology, learns from them, being able to identify small elements that initially presented shortcomings in a single pass after training. It can be ascertained that the metrics achieved by the presented proposal surpass the original model. Selecting sequence number two as an example, an increase from an average mAP of 27.7% with the raw model to 92.6% after applying the proposal is obtained. This trend is consistent across the remaining sequences.

Following this line, the work published in the IWANN (International Work Conference on Artificial Neural Networks) conference (chapter 7) also focuses on applying techniques based on re-referencing super-resolved areas. However, the methodology is modified to detect the areas of interest, which in this case is the detection of roads and lanes through binarization. Once identified, super-resolved areas are generated exclusively in these zones optimally through a previously established heuristic. The results show that applying this new methodology significantly reduces the number of generated areas and improves the accuracy of the object detection model. The presented methodology yields a significant improvement in mAP compared to the RAW model, with an average increase of up to 22.7% across the evaluated test sequences. Furthermore, the presented approach detects a more substantial number of objects, encompassing small and medium-sized objects, without requiring the retraining of the object detection model. However, certain cases may require a balance between the number of regions generated and the accuracy achieved by the model. Thus, this line of work culminates in the publication of Neural Computing and Applications (chapter 8). The methodology uses the same concepts established in previous work. However, it redefines the generation of zones to apply SR and re-inference. A balance between processing time and model accuracy is achieved by defining the size of the sliding windows together with the Bron-Kerbosch algorithm for maximum clique generation. A clique is a set of

elements, each of which is a vehicle in traffic sequences. The results support this methodology, which achieves better accuracy while reducing processing time, showcasing an increase of up to 44.6% in detection rates. This transition is observed from an average detection rate of 14.5% for the EfficientDet D4 model to 59.1% when employing the methodology presented for the first sequence.

In parallel but closely related, these methodologies described above are applied in object segmentation, with the second group conforming based on the works presented. The first of these was presented in IWINAC (International Work-Conference on the Interplay Between Natural and Artificial Computation) (chapter 4). It proposes a super-resolution re-referencing method like the one shown in chapter 3. However, it presents the peculiarity of using masks instead of bounding boxes. This allows a more precise identification of the element. The significant improvement achieved by the approach is demonstrated, establishing its potential through a series of experiments. For example, the Person class increased from an average score of 89.1% to 93.5%. Additionally, the item detection rate improved from 42.71% to 71.53%. As a successor of the advances developed in this work, the methodology presented in ICAE (Integrated Computer-Aided Engineering) (chapter 5) is presented, where again an optimization of the zones to be re-inferred is performed, and a final process is applied to improve the masks of the elements identified multiple times. The results obtained with different object segmentation models support the accuracy improvement after applying the described methodology, revealing improvements of up to 8.1% when implemented with the YOLACT++ model in one of the sequences.

Finally, the third group of papers focuses on anomaly detection. The work presented in Sensors (chapter 9) uses a methodology similar to that described in work such as that presented in chapter 7 for improved feature identification. However, to subsequently identify elements located in unusual regions, the common regions of each class are computed. For example, the proposed approach employing the EfficientDet D4 model demonstrated an enhancement in mAP scores for the pedestrian class in Sequence 1, achieving a value of 12.8% compared to the base model's score of 1.8%.

These papers have approached the problem of small object detection, segmentation, and anomaly identification in different ways, using similar tools and paradigms to solve them.

10.2 Future Work

According to the research conducted in the elaboration of this thesis, a series of conclusions and future lines of research are established to continue the developments achieved. Therefore, the ways to improve the presented proposal are determined, as well as the formulation of new ideas to be applied in new areas. The problem of detection and segmentation of small objects must be highlighted as a conclusion of this thesis. While DL has achived significant advancements, tasks still require further improvement. The detection and segmentation of this type of element needs to be improved in many cases, constituting a **complex task**. In this thesis, a series of works are carried out that try to **mitigate** this problem, obtaining good results. Therefore, it may be particularly attractive to continue working in this direction to provide more robust solutions.

The future lines are divided into three main groups, according to the works that support this thesis:

- Continue to work on object recognition models to improve model accuracy further.
- Continue in the area of segmentation by providing new techniques to improve the accuracy of the masks produced by the model.
- Extend the study of anomaly detection to new areas, such as trajectory anomalies in video sequences.

10.2.1 Improving Small Object Detection

The first line of work, which is the most feasible in the short term, is based on developing and improving the methodology developed in the work proposed throughout the thesis to improve the accuracy rate. Therefore, there are three points where the possible improvements to be made are defined:

- Ensemble object recognition: A wide range of object detection models with different architectures could be experimented with. The goal would be to combine the results of each model to create a unified solution that provides robustness.
- Feature Selection: Investigate and apply other feature selection algorithms to improve object localization accuracy in static scenes. Such algorithms could be used to identify more descriptive features for subsequent detection of objects in these regions.
- Temporal information integration: Developing techniques for integrating temporal information into object detection in static video sequences. Some of these techniques would correspond to Kalman filters.
- Applying Other Image Quality Enhancement Techniques: Advanced image enhancement techniques such as noise reduction, contrast enhancement, and color correction could improve image quality beyond SR. This could improve object detection accuracy and scene analysis.

10.2.2 Improving Small Object Segmentation

The second line, based on improving the segmentation of small objects, is again one of the most viable ways to continue developing new techniques that can improve the proposed methodology. For this purpose, the following possible improvements are identified:

- Advanced Feature Selection Techniques: Explore and implement the proposed methodology advanced methods based on feature selection beyond SIFT (Scale Invariant Feature Transform) Lowe (2004) or SURF (Speeded Up Robust Features) Bay et al. (2006), among others, to prioritize relevant features in the scene and determine the areas to focus on for more accurate segmentation.

- Data augmentation strategies and transfer learning: The application of different data augmentation techniques adapted to segmentation tasks, such as a series of transformations simulating certain conditions or the generation of synthetic images using GAN (Generative Adversarial Networks) generative networks according to the input data, can lead to better training to perform transfer learning with previously trained models to improve object segmentation. The analysis could be performed based on model parameter tuning to determine which model performs best in a range of domains.

10.2.3 Anomaly Detection

Although the scope of the thesis has focused on improving the detection and segmentation of small objects, parts of this methodology have also been applied to identifying anomalies. Therefore, the proposed lines for detecting anomalies in the location of objects in unusual areas are as follows:

- Improve robustness under challenging lighting conditions: Further investigate and develop techniques to improve the methodology's performance in adverse lighting conditions. This could include developing advanced illumination correction methods beyond LUT (Look-Up Tables), such as deep learning-based approaches or adaptive exposure control.
- Real-world testing and validation: Conduct extensive real-world testing and validation in various surveillance environments, including urban, rural, and remote locations. Collect data under different lighting conditions and scenarios to assess the reliability and effectiveness of the methodology.

Appendix A

Resumen de publicaciones obtenidas

Resumen:

En este apéndice se muestran una lista de tablas que resume la información asociada a los trabajos publicados que conforman esta tesis. Para las revistas, se ha considerado el ranking *Journal Citation Reports (JCR)* del año correspondiente a la publicación, mientras que para los congresos se indica el ranking CORE [1] del año correspondiente, o en su defecto, el último en el caso de los publicados en 2022 pues estos aún no han sido establecidos y el *GII-GRIN-SCIE (GGS) Conference Rating*[2] del año 2021.

Título	Improved detection of small objects in road network sequences using CNN and super resolution
Autores	Iván García-Aguilar, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Expert Systems
Año	2021
Factor de impacto	2.812
Categorías JCR	COMPUTER SCIENCE, THEORY & METHODS (37/110) (Q2)
Estado	Publicado
DOI	`https://doi.org/10.1111/exsy.12930`
Referencia	García-Aguilar et al. 2022

[1] `https://www.core.edu.au/`
[2] `https://scie.lcc.uma.es:8443/`

Título	Enhanced Image Segmentation by a Novel Test Time Augmentation and Super-Resolution
Autores	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio, Enrique Domínguez
Congreso	9th INTERNATIONAL WORK-CONFERENCE ON THE INTERPLAY BETWEEN NATURAL AND ARTIFICIAL COMPUTATION (IWINAC)
Año	2022
GGS Rating	Work in Progress
CORE Rating	-
Estado	Publicado
DOI	https://doi.org/10.1007/978-3-031-06527-9_15
Referencia	García-Aguilar et al. 2022

Título	Optimized instance segmentation by super-resolution and maximal clique generation
Autores	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio, Enrique Domínguez
Revista	Integrated Computer-Aided Engineering (ICAE)
Año	2023
Factor de impacto	6.5
Categorías JCR	COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (22/110) (Q1))
Estado	Publicado
DOI	https://doi.org/10.3233/ica-230700
Referencia	García-Aguilar et al. 2023b

Título	Automatic labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks
Autores	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Pattern Recognition Letters (PRL)
Año	2023
Factor de impacto	5.1
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (52/145) (Q2))
Estado	Publicado
DOI	https://doi.org/10.1016/j.patrec.2023.01.015
Referencia	García-Aguilar et al. 2023

Título	Minimal Optimal Region Generation for Enhanced Object Detection in Aerial Images using Super-Resolution and Convolutional Neural Networks
Autores	Iván García-Aguilar, Lipika Deka, Rafael M. Luque-Baena, Enrique Domínguez, Ezequiel López-Rubio
Congreso	17th International Work Conference on Artificial Neural Networks (IWANN)
Año	2023
GGS Rating	Work in Progress
CORE Rating	-
Estado	Publicado
DOI	https://doi.org/10.1007/978-3-031-43085-5_22
Referencia	García-Aguilar et al. 2023

Título	Object detection in traffic videos: an optimized approach using super-resolution and maximal clique algorithm
Autores	Iván García-Aguilar, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Neural Computing and Applications
Año	2023
Factor de impacto	6.0
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (41/145) (Q2))
Estado	Publicado
DOI	https://doi.org/10.1007/s00521-023-08741-4
Referencia	García-Aguilar et al. 2023a

Título	Small-Scale Urban Object Anomaly Detection Using Convolutional Neural Networks with Probability Estimation
Autores	Iván García-Aguilar, Rafael M. Luque-Baena, Enrique Domínguez, Ezequiel López-Rubio
Revista	SENSORS
Año	2023
Factor de impacto	3.9
Categorías JCR	ENGINEERING, ELECTRICAL & ELECTRONIC (100/275) (Q2))
Estado	Publicado
DOI	https://doi.org/10.3390/s23167185
Referencia	García-Aguilar et al. 2023

Appendix B

Resumen en Español

En la actualidad, la vertiginosa proliferación de sistemas orientados a contenidos multimedia ha provocado un aumento de los datos recogidos, permitiendo diversos avances en inteligencia artificial y visión por computador en la última década. Dado este aumento, analizar los mismos de forma manual constituye un reto, pues es difícil afrontar dicha tarea en base al tiempo que se requiere. Además, el uso de métodos clásicos presentan una serie de deficiencias que deben ser resueltas. Se necesitan por ello nuevas técnicas para mitigar este problema. El campo del aprendizaje profundo basado en redes neuronales convolucionales es la solución para superar estas limitaciones.

El principal problema abordado en esta tesis doctoral es la mejora de la detección de objetos pequeños en secuencias de vídeo de carretera utilizando redes neuronales convolucionales y super-resolución. El objetivo es incrementar significativamente el número de píxeles que conforma la imagen de entrada para así mejorar la inferencia y el procesamiento por parte de los modelos de detección de objetos. Como problema derivado, se aborda la detección de anomalías utilizando estas herramientas. El enfoque basado en la mejora de la detección de objetos pequeños en secuencias de vídeo de carreteras constituye un área de investigación crucial debido a múltiples razones. Avances actuales en el campo de la detección de objetos como los modelos entrenados en ImageNet han demostrado un buen rendimiento a la hora de identificar objetos grandes y prominentes. Sin embargo, aún enfrentan una serie de desafíos a solventar en el ámbito de los objetos pequeños, los cuales son comúnes en escenarios de tráfico. Por ello, esta línea de investigación representa un problema relevante y desafiante dentro del campo.

Como testigo de estos objetivos, se incluyen una serie de trabajos de investigación realizados durante la tesis. El primero fue publicado en la revista Expert Systems en 2021 y trata sobre la mejora de la detección de objetos pequeños utilizando técnicas de super-resolución y re-inferencia para mejorar la precisión obtenida por los modelos de detección de objetos. La propuesta mejora el rendimiento con respecto a modelos pre-entranados, consiguiendo una precisión media del 45.1% con el modelo EfficientDet-D4 para la primera secuencia de vídeo, una mejora significativa del 24.3% con respecto al modelo base. El segundo trabajo se presentó en la *International Work Conference on the Interplay Between Natural and Artificial Computation* en 2022 y propone aplicar estas técnicas en entornos de segmentación de objetos aplicados a secuencias urbanas. La propuesta presentada demuestra

una notable mejora en multitud de objetos. Por ejemplo, la clase Persona pasa de un 89.1% a un 93.5%. Además, la tasa de detección de objetos mejora de un 42.71% a un 71.53%. El tercer trabajo, publicado en la revista *Integrated Computer-Aided Engineering* en 2023, trata sobre la aplicación de las técnicas anteriormente descritas en la segmentación de objetos en secuencias urbanas, utilizando heurísticas para optimizar el rendimiento de la metodología establecida. Los experimentos respaldan la efectividad de la propuesta presentada, mostrando una mejora de hasta un 8.1% con el modelo YOLACT++ para una de las secuencias, mostrando un beneficio en cuanto a la segmentación en sistemas de video-vigilancia se refiere. El cuarto trabajo se publicó en *Pattern Recognition Letters* en 2023 y se basa en el desarrollo de una metodología que permite el entrenamiento automático mediante el ajuste fino de un modelo de reconocimiento de objetos sin anotaciones manuales para mejorar la identificación de elementos pequeños. Las métricas obtenidas con la propuesta presentada logran sobrepasar al modelo original. Por ejemplo, en la secuencia dos, la tasa de precisión media aumenta de un 27.7% a un 92.6%, mejorando por ello la precisión en multitud de dominios. El quinto trabajo se presentó en la *International Work Conference on Artificial Neural Networks* en 2023 y se basa en la generación mínima de zonas de interés en una secuencia en función del cálculo de la densidad de los elementos necesarios a partir de una matriz computada. La metodología presentada mejora significativamente hasta un 22.7% comparado con el modelo original en las múltiples secuencias evaluadas, detectando por ello más objetos sin necesidad de re-entrenar el modelo. El sexto trabajo fue presentado en la revista *Neural Computing and Applications* en 2023 y trata sobre la implementación de técnicas de optimización mediante un grafo computado, junto con técnicas de super-resolución para reducir relativamente el número de sub-imágenes sobre las que re-inferir, mejorando así significativamente el tiempo de procesado. Los resultados reflejan un aumento en la tasa de detección del 44.6%, pasando de un 14.5% a un 59.1% con el modelo EfficientDet D4 usando la propuesta presentada en la primera secuencia. Por último, el séptimo trabajo de esta tesis fue publicado en *Sensors* en 2023, donde se presenta una metodología para la detección de anomalías y amenazas en secuencias sintéticas de peatones. Esta metodología incluye técnicas de mejora de la imagen para aumentar las detecciones obtenidas por la red neuronal seleccionada.

Estos siete trabajos constituyen el memorándum de esta tesis doctoral y el autor presenta los resultados en estos años de investigación.

B.1 Introducción

El auge de la tecnología ha supuesto una revolución que ha provocado un cambio notorio en la sociedad. Entre otras cosas, la accesibilidad y el coste reducido de los dispositivos y los distintos servicios han promovido estos aspectos, suponiendo un cambio en cuanto a nuestra forma de vivir y trabajar. Hace unos años, los ordenadores personales o el acceso a Internet suponían un coste elevado que sólo unos pocos podían permitirse. Hoy, la reducción del coste de estos dispositivos los ha hecho más accesibles. Como consecuencia, la cantidad de información y contenidos disponibles ha aumentado considerablemente.

La reducción del coste de estos sistemas no está directamente relacionada con los ordenadores, sino que va más allá de estos. Este hecho también afecta al resto de

dispositivos electrónicos, en los que se ha ido mejorando con el paso de los años, reduciendo así su coste. Entre estos dispositivos, por ejemplo, hay sistemas centrados en la recogida de imágenes y vídeos, como las cámaras de video-vigilancia. La información recogida es fácilmente accesible a través de diversas plataformas en la red. Hoy en día, muchos sistemas están especializados en diferentes tareas relacionadas con la recopilación de contenidos multimedia. Los más comunes son las cámaras 2D, que recogen fotos o secuencias de vídeo para su posterior procesamiento. Algunos ejemplos podrían ser la cámara que poseen dispositivos inteligentes como los *smartphones* o los sistemas de video-vigilancia, los cuales suelen colocarse en puntos elevados para el control del tráfico entre otros. También existen sistemas integrados en los vehículos, denominados *dashcams*, los cuales capturan la carretera desde el punto de vista del conductor.

Los datos recogidos por estos dispositivos pueden tener fines personales y profesionales. En este último grupo se incluye la recogida de datos para uso privado, ya sea para empresas o entidades públicas, como el gobierno, que tienen fines específicos. Algunos ejemplos son la instalación de cámaras y sistemas de video-vigilancia, por ejemplo en las calles de las ciudades, para controlar la aglomeración de personas en un área concreta, o en zonas de carretera para controlar el tráfico e identificar posibles peligros que puedan surgir.

Con el auge de este tipo de sistemas, se produce un aumento exponencial en la recopilación de datos, lo que plantea varios retos y problemas que son necesario abordar. En muchos casos, como el uso personal de estos dispositivos, la información recopilada es prescindible y se utiliza únicamente con fines de entretenimiento. Sin embargo, hay otros ámbitos en los que el análisis de la información recopilada es esencial, ya que de lo contrario, no tendría sentido recopilar esto. En el ámbito de la gestión y el control de carreteras, los sistemas de video-vigilancia captan imágenes las 24 horas del día. Analizar las imágenes captadas por estos sistemas supone un reto. Una sola grabación que cubra todo un día requiere el análisis por parte de varias personas designadas a tal efecto. Por ello, es necesario aportar nuevos métodos y avances para analizar estas secuencias de forma óptima y eficaz.

En los últimos años, se han desarrollado numerosos avances en el campo de la visión por computador, los cuales permiten procesar imágenes y vídeos según los requisitos establecidos. Por otro lado, gracias al incremento de estos datos, el campo del aprendizaje profundo se ha visto influenciado de forma significativa, propiciando el desarrollo de modelos que aprenden de esta información. Todo ello ha dado lugar a un conjunto de modelos tanto de detección como de segmentación que permiten identificar una amplia gama de elementos. Sin embargo, siguen presentando deficiencias a la hora de identificar elementos pequeños. Su precisión media (mAP) disminuye significativamente cuando el tamaño de estos elementos es pequeño. Todo ello se debe en parte al comportamiento intrínseco del propio modelo. Al realizar la inferencia, se reduce el número de píxeles de la imagen dada como entrada, haciendo que aquellos elementos compuestos por un bajo número de píxeles se pierdan en el procesado. Gracias a las propuestas presentadas en esta tesis, ha sido posible mejorar la tasa de precisión media de una amplia gama de modelos de detección y segmentación de objetos sin necesidad de modificar la estructura interna de los mismos. Se aplican técnicas de aumentado de datos y re-inferencia, proporcionando así una metodología que permite mejorar la identificación de este tipo de elementos en diferentes secuencias, principalmente centradas en entornos de carretera.

B.2 Estado del Arte

En el capítulo 2 se detallan los fundamentos teóricos que respaldan los trabajos presentados en esta tesis, incluyendo por ello los problemas relacionados con la detección y segmentación de objetos pequeños, así como la la detección de anomalías así, determinando por ello una serie de aproximaciones para resolverlos.

Para comenzar, la sección 2.1 describe los conceptos fundamentales que sientan las bases para comprender correctamente el funcionamiento de las redes neuronales. En primer lugar, se inicia con una introducción al concepto más básico y elemental correspondiente a la neurona artificial, detallando sus limitaciones. Posteriormente, se profundiza en las capas que componen las redes neuronales, describiendo por ello las capas convolucionales, de activación, agrupación, densas, normalización por lotes y capas de eliminación aleatoria. Seguido de ello se describe la función de pérdida, destacando su papel en la valiación del rendimiento en los modelos. Finalmente, se exploran las diversas estrategias de entrenamiento existentes, incluyendo por ello el entrenamiento desde cero, ajuste fino y aprendizaje por transferencia. De este modo, se proporciona una comprensión completa de como se entrenan estos modelos basados en redes neuronales.

En la sección 2.2 se detallan los fundamentos de la super–resolución, un campo crucial en el contexto de la visión por computador. Se comienza explorando las diversas técnicas clásicas centradas en la mejora de calidad de la imagen. A continuación, se profundiza los fundamentos de la super-resolución, destacando los principios y bases teóricas que conforman este punto. Finalmente, se examinan los modelos específicos empleados para llevar a cabo este proceso, otorgando por ello una visión general de las diversas aproximaciones disponibles.

Seguidamente, la sección 2.3 detalla el problema de la detección y segmentación de objetos pequeños. Por ello, se establecen previamente los conceptos básicos para ambos campos, detallando las diferencias entre reconocimiento, detección y segmentación de objetos. Dentro del apartado de detección, se realiza una división en base al funcionamiento de los modelos, realizando una descripción general del funcionamiento, limitaciones y beneficios que presentan cada uno de ellos. Acto seguido, se sigue la misma estructura dentro del campo de la segmentación de objetos, especificando por ello la diferencia entre segmentación a nivel de píxel, semántica y de instancias. Para estos dos últimos grupos, se enuncian los modelos más conocidos. Estos conceptos conforman las bases para entender las investigaciones y trabajos presentados en esta tesis. Tras ello, se detalla formalmente el problema, estableciendo además las posibles soluciones para mitigar el mismo, los retos a los que hacer frente, así como las aplicaciones que pueden beneficiarse de las soluciones aportadas. Dentro de estos problemas, se detallan aspectos como la variedad de tamaños que presentan los objetos en las imágenes, los entornos complejos, así como aspectos relacionados con la iluminación, calidad de imagen o oclusión entre otros.

Dado que es el problema principal de la tesis, se establecen una serie de propuestas y trabajos previos para solucionar el mismo. En sus primeras etapas, se establecen una serie de avances centrados en las características de los elementos superficiales y los propios clasificadores, destacando trabajos como el propuesto por Kembhavi et al. (2011) donde se capturan características estructurales de los objetos a través de mapas de probabilidad de calor, historigramas de gradientes orientados y un descriptor de imagen. Otros avances se centraron en colores y formas

de los elementos a detectar, como es el caso de Le et al. (2010). El avance de las redes neuronales convolucionales ha dado lugar a métodos centrados en el aprendizaje profundo, surgiendo por ello multitud de métodos de detección de objetos con trabajos pioneros como Girshick et al. (2014); Girshick (2015), en los cuales se proponen algoritmos simples y escalables que mejoran la precisión mediante la aplicación conjunta de propuestas regionales junto con redes profundas, dando lugar así a modelos más sofisticados como los propuestos por Ren et al. (2017); Liu et al. (2016); Tan and Le (2019); Zhu et al. (2021). La segmentación de objetos surge como una evolución natural de la detección de objetos, dando lugar así a una serie de modelos He et al. (2017b); Hayder et al. (2017).

Sin embargo, en el campo de la detección de objetos pequeños, dado que es un problema relativamente reciente, existen pocos estudios centrados en resolver el mismo. Chen et al. (2017) establecieron un conjunto de datos y métricas específicas para la detección de objetos pequeños, marcando un punto de referencia en el campo. Krishna and Jawahar (2017) mejoraron esto mediante una técnica de muestreo ascendente. Los métodos para la detección de objetos pequeños se han centrado en crear arquitecturas específicas y técnicas de mejora, como el enfoque multi-escala de Lin et al. (2017), que presenta la red piramidal de características (FPN) para mejorar la detección de objetos. Existen otros enfoques basados en técnicas de aumento de datos. Kisantal et al. (2019) sobre-muestrean imágenes con objetos pequeños, copiándolos y pegándolos varias veces para mejorar la calidad del modelo de detección. Akyon et al. (2022) proponen SAHI (Slicing Aided Hyper Inference), una solución que utiliza la segmentación para mejorar el rendimiento, manteniendo la complejidad y los requisitos computacionales bajos. Además, en el ámbito de la segmentación, se han desarrollado estrategias que aprovechan datos etiquetados a nivel de imagen, como la generación de pseudo-máscaras Laradji et al. (2019), o incluso la anotación de un solo píxel por objeto Laradji et al. (2020). Otras propuestas se centran en mejorar el rendimiento de las redes profundas, como la estructura de propuesta de región multi-escala de Yang et al. (2020) o la red MR R-CNN propuesta por Zhang et al. (2020), que analiza la relación entre la información espacial y el tamaño del objeto. También se ha utilizado la generación de datos artificiales, como lo hacen Ghiasi et al. (2021) al generar datos sintéticos mediante el recorte y pegado de instancias de objetos en otras imágenes. El uso de redes generativas adversiales (GAN) han supuesto una vía de desarrollo, dando lugar a trabajos como el propuesto por Bai et al. (2018), donde se presenta un enfoque llamado Generative Adversarial Network (MTGAN). El generador se encarga de mejorar la calidad de las imágenes, y el discriminador realiza tareas de clasificación y detección.

Seguido de esto, la sección 2.4 detalla uno de los temas secundarios de la tesis, basada en la detección de anomalías. Por ello, de nuevo se presenta el problema, así como las posibles soluciones a realizar para mitigar el mismo. Se han desarrollado diversas técnicas para la detección de anomalías en secuencias de vídeo. Inicialmente, se utilizaron métodos clásicos, sin embargo, poseían una serie de limitaciones a la hora de capturar distribuciones complejas. El uso de las redes neuronales convolucionales ha impulsado la investigación en este campo. Hu et al. (2020) propusieron un marco de aprendizaje débilmente supervisado para identificar comportamientos inusuales en escenas. Kanu-Asiegbu et al. (2021) presentó un enfoque innovador para detectar eventos anómalos en peatones utilizando errores de predicción de trayectorias. Otros estudios han empleado YOLOv3 para la detección

de peatones y redes LSTM para identificar comportamientos anómalos tales como el trabajo propuesto por Chang et al. (2022), así como algoritmos de optimización y redes neuronales convolucionales profundas en base a trabajos como el de Gayal and Patil (2023). Además, se ha utilizado la generación de datos sintéticos a través del simulador CARLA Dosovitskiy et al. (2017). B. Sophia (2023) proponen un modelo PFPN-ADT para pasos de peatones. Se ha trabajado adicionalmente en la detección de anomalías en comportamientos de peatones, utilizando características de movimiento-apariencia y técnicas de aprendizaje profundo, como DLADT-PW Pustokhina et al. (2021). Estos avances buscan mejorar la seguridad a través del análisis inteligente de vídeos, sin embargo, presentan de nuevo deficiencias en ciertos contextos.

B.3 Trabajos que apoyan esta Tesis

El **primero** de los siete trabajos que avalan esta tesis se denomina *Improved detection of small objects in road network sequences using CNN and super resolution* (*Mejora en la detección de objetos pequeños en secuencias de carretera utilizando redes neuronales convolucionales y super-resolución* en castellano). Fue publicado en el año 2021 en la revista *Expert Systems*, ocupando en dicho año la posición del segundo cuartil (37/110) de la clasificación JCR en la categoría Ciencia Computacional, Teoría y Métodos. Dicho trabajo se relaciona con la mejora en la detección de objetos pequeños mediante la re-inferencia en zonas super-resueltas. Por ello, se plantea un metamétodo y una serie de experimentos con resultados que avalan la aplicación del mismo.

El funcionamiento del meta-método es el siguiente. Partiendo de un modelo de detección de objetos pre-entrenado, denotado como $\mathcal{F}$, toma una imagen de entrada de baja resolución $\mathbf{X}LR$ con el fin de inferir sobre la misma y producir un conjunto de detecciones iniciales S. Cada detección S_i en S se representa a través de las coordenadas donde se localiza el cuadro delimitador de cada objeto, siendo estas (a_i, b_i) y (c_i, d_i) para las esquinas superior izquierda e inferior derecha. Además, se almacena la clase a la que pertenece q_i, y una puntuación de confianza r_i. Posteriormente, se aplica super-resolución a la imagen de entrada haciendo uso de un modelo pre-entrenado $\mathcal{G}$. Tras este proceso, se obtiene una versión de alta resolución, denominada como $\mathbf{X}HR$. Para refinar aún más la imagen, se emplea un procedimiento de eliminación de ruido $\mathcal{D}$ en $\mathbf{X}HR$. Esta imagen se utiliza para generar sub-imágenes alrededor de cada detección inicial. Cada una de ellas corresponderá con el área alrededor de una detección. Se re-infiere de nuevo sobre estas zonas, dando lugar así a nuevas detecciones S_i. Estas nuevas detecciones se traducen en el espacio de coordenadas de la imagen original y se aplica un proceso de filtrado utilizando la métrica de Intersección sobre Unión (IOU) para eliminar detecciones duplicadas, dando lugar así a un conjunto mejorado de detecciones con mayor precisión.

Los experimentos se realizan sobre diversas secuencias que han sido anotadas manualmente a partir del conjunto de datos NGSIM U.S. Department Of Transportation Federal Highway Administration (2017). En total, se evalúan tres secuencias cuantitativamente. Se aportan resultados en imágenes de otros datasets como el de M-30-HD Guerrero-Gómez-Olmedo et al. (2013) y UA-DETRAC Wen et al. (2020); Lyu et al. (2018, 2017). Las secuencias son analizadas aplicando la

metodología propuesta a diversos modelos de detección de objetos de cara a determinar si la misma supone una solución efectiva. Como métricas, se analiza la tasa de precisión media (mAP) obtenida.

Los resultados obtenidos muestran que la metodología mejora la precisión media en cada uno de los modelos evaluados. Alguno de ellos tales como *EfficientDet D3*, aumenta su *mAP* en un 17,6% en promedio, pasando de una precisión del 9,1% al 29,7%. Esta mejora también se aplica a elementos medianos y grandes, obteniendo una mejora general del 17,1% para este modelo, aumentando así el número de elementos detectados en la imagen de entrada y mejorando la puntuación de clase inferida sin necesidad de modificar la arquitectura de dicho modelo o re-entrenar el mismo.

El **segundo** de los trabajos presentados se denomina *Enhanced Image Segmentation by a Novel Test Time Augmentation and Super-Resolution* (*Mejora en la segmentación de imágenes mediante aumento de datos y super-resolución*) y fue publicado en el año 2022 en la *9th International Work-Conference on the Interplay Between Natural and Artificial Computation IWINAC*. Dicho trabajo aplica la metodología descrita anteriormente en el ámbito de la segmentación de objetos. Para ello, se evalúa la misma haciendo uso del modelo conocido como Mask R-CNN. En lugar de trabajar con cuadros delimitadores, ahora se hace uso de las máscaras obtenidas para cada elemento.

Los experimentos se realizan sobre secuencias que conforman el conjunto de datos de CityScapes Cordts et al. (2016) y Visdrone Zhu et al. (2022). Se obtiene una mejora significativa en la precisión. Por ejemplo, en la secuencia capturada en *Frankfurt* se pasa de tener una precisión del 13% al 20.2%. Además, se consigue un incremento en la confiabilidad de las detecciones. La clase persona pasa de una puntuación media del 89.1% al 93.5%, mejorando además el número de elementos detectados de un 42.71% hasta un 71.53%.

El **tercer** trabajo presentado lleva por título *Optimized instance segmentation by super-resolution and maximal clique generation* (*Segmentación de instancias optimizada mediante super-resolución y generación de cliques máximos*). El artículo se publicó en la revista ICAE en 2023 y en dicho año ocupaba posición en el primer cuartil (22/110) en la categoría Ciencia Computacional, Aplicaciones Interdisciplinares de la clasificación JCR.

El sistema propuesto para mejorar la segmentación de instancias consta de varios módulos. En el módulo inicial, se utiliza una red neuronal profunda $\mathcal{F}$ para segmentar objetos en una imagen de entrada $\mathbf{X}$, generando un conjunto de detecciones S. Posteriormente, se aplica super-resolución a la imagen para obtener una versión de alta resolución, lo que aumenta la calidad de las detecciones. Acto seguido, se calcula un grafo que conecta las detecciones iniciales, permitiendo identificar qué pares de detecciones podrían beneficiarse de una ejecución conjunta de la red de segmentación en la misma ventana de alta resolución, lo que reduce significativamente la carga computacional. Se utiliza el algoritmo de Bron-Kerbosch para identificar cliques en el grafo, que son conjuntos de detecciones completamente conectadas entre sí. Luego, se aplica un algoritmo de selección basado en la cantidad de nodos en cada clique y la frecuencia con la que se utilizan. Esto permite elegir un conjunto reducido de cliques, lo que a su vez se traduce en procesar un conjunto más pequeño de ventanas de alta resolución.

Como experimentos, se aplica la metodología descrita a diversos modelos orientados a la segmentación de objetos. Para ello, se utiliza el conjunto de datos

CityScapes Cordts et al. (2016). Los resultados experimentales muestran que el enfoque mejora significativamente la precisión del modelo base de segmentación de imágenes. Específicamente, la propuesta detecta más objetos con una mayor precisión media en una variedad de secuencias del mundo real sin modificar las capas que componen el modelo original ni volver a entrenarlo. Modelos como YOLACT++ logran una ganancia máxima de hasta un 8.1%, aumentando el número promedio de elementos detectados y reduciendo la cantidad de sub-imágenes en las que el modelo debe volver a inferir.

El **cuarto** trabajo referido es una metodología titulada *Automatic labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks* (*Etiquetado automático de datos de entrenamiento para una mejora en la detección de objetos en vídeos de tráfico mediante redes neuronales convolucionales profundas re-entrenadas*). El artículo se publicó en la revista PRL (Pattern Recognition Letters) en 2023 y en dicho año ocupaba posición en el segundo cuartil (52/145) en la categoría inteligencia artificial en la clasificación del JCR.

La metodología propuesta para mejorar el rendimiento de las detecciones de objetos mediante redes neuronales convolucionales profundas se compone de dos subsistemas. En primer lugar, se refina la salida de una red neuronal convolucional para la detección de objetos mediante super-resolución, lo que genera un conjunto de detecciones de alta calidad y un conjunto de datos etiquetado automáticamente. Este subsistema involucra la conversión de un video sin etiquetar en un conjunto de imágenes que se procesan en la etapa siguiente. Luego, las detecciones de alta calidad se utilizan para construir un nuevo conjunto de entrenamiento para la red de detección de objetos con el fin de ajustar finamente el modelo y mejorar su rendimiento. Esta técnica mejora la precisión del modelo sin modificar sus capas originales ni volver a entrenarlo.

Los resultados experimentales sobre el conjunto NGSIM U.S. Department Of Transportation Federal Highway Administration (2017) demuestran un aumento significativo en la precisión de la detección de objetos en una gran variedad de secuencias del mundo real sin requerir modificaciones significativas en el modelo original.

El **quinto** trabajo de esta tesis se denomina *Minimal Optimal Region Generation for Enhanced Object Detection in Aerial Images using Super-Resolution and Convolutional Neural Networks* (*Generación Mínima de Regiones óptimas para una Detección Mejorada de Objetos en Imágenes Aéreas mediante el Uso de Super-Resolución y Redes Neuronales Convolucionales.*). Fue publicado en el año 2023 en la *17th International Work Conference on Artificial Neural Networks IWANN*. Este artículo presenta una metodología para mejorar la detección de vehículos en secuencias de imágenes aéreas, comenzando con una secuencia de video no etiquetada, $\mathbf{D}$, compuesta por un conjunto de imágenes ($\mathbf{Y}_l$), donde l varía de 1 a N. Cada imagen se procesa mediante binarización basada en saturación en el espacio de color HSV para obtener una máscara binaria, y se aplica un modelo de detección de objetos YOLOv7, representado por $\mathcal{G}$, para obtener un conjunto de regiones de interés ROI, que se almacenan en $\mathbf{W}$. Posteriormente, se combina la información de binarización y las ROIs generadas por YOLOv7 para calcular un conjunto mínimo óptimo de ventanas que cubran un umbral mínimo de cobertura. Esto se logra mediante la generación de todas las cajas delimitadoras posibles y el cálculo de la cantidad de píxeles dentro de cada caja. Después de la normalización,

se calcula una matriz ponderada denominada **computed_grid** que prioriza las ROIs propuestas por YOLOv7. A continuación, las ROIs seleccionadas se someten a super-resolución utilizando el modelo Real-ESRGAN, que mejora la calidad de estas regiones. Las regiones super-resueltas se someten nuevamente al modelo YOLOv7 para obtener un nuevo conjunto de detecciones W_{HR}. Las ubicaciones de estas detecciones se traducen del sistema de coordenadas de las imágenes super-resueltas al de las imágenes originales. Finalmente, se evalúan las mismas para determinar si corresponden a objetos previamente no detectados o a detecciones existentes mediante la métrica de intersección sobre unión (IOU). Las detecciones con un IOU por encima de un umbral θ se consideran del mismo objeto. Por cada elemento, se selecciona la detección con la puntuación más alta.

Para ello, se evalúa en una serie de secuencias del conjunto de datos UAVDT Du et al. (2018). Los resultados del estudio realizado demuestra que este enfoque mejora de manera efectiva la precisión de la detección de objetos en una variedad de escenarios. Específicamente, la metodología mejora significativamente el mAP en comparación con el modelo original, con un aumento promedio de hasta un 22.7% en las secuencias de prueba evaluadas. Además, se detecta un mayor número de objetos, incluyendo objetos pequeños y medianos, sin necesidad de volver a entrenar el modelo de detección de objetos.

El **sexto** trabajo referido es una metodología titulada *Object detection in traffic videos: an optimized approach using super-resolution and maximal clique algorithm* (*Detección de objetos en videos de tráfico: un enfoque optimizado utilizando super-resolución y algoritmo de clique máximo.*). El artículo se publicó en la revista *Neural Computing and Applications* en 2023 y en dicho año ocupaba posición en el segundo cuartil (41/145) en la categoría inteligencia artificial de la clasificación del JCR. Este trabajo combina técnicas de optimización y super-resolución para reducir el número de áreas que deben ser inferidas utilizando un modelo de detección de objetos, mejorando así significativamente el tiempo de ejecución.

En esta sección se presenta una metodología que aprovecha modelos de detección basados en redes neuronales convolucionales para el procesamiento de alta resolución. Se comienza aplicando una red de super-resolución $\mathcal{G}$ a imágenes de baja resolución para obtener versiones de alta resolución. Luego, se utiliza un modelo $\mathcal{F}$ para realizar detecciones tentativas en las imágenes de baja resolución. Estas detecciones se filtran mediante un umbral de confianza T y se obtiene un conjunto de detecciones de alta confianza. Se genera un grafo que conecta estas detecciones, y se buscan cliques máximos para agrupar detecciones relacionadas. Un algoritmo voraz selecciona un conjunto de cliques que cubren todas las detecciones. A partir de estos cliques, se generan sub-imágenes óptimas para una segunda inferencia de detección, lo que reduce el tiempo de ejecución. Finalmente, se agrupan las detecciones basadas en la medida de Intersección sobre Unión para obtener una mayor cantidad de detecciones con una mejor inferencia de clase. Por ello, sigue los mismos principios que el tercer trabajo publicado pero en el contexto de la detección de objetos.

Para ello, se realizan una serie de experimentos con secuencias del conjunto de datos del NGSIM U.S. Department Of Transportation Federal Highway Administration (2017) y VisDrone Zhu et al. (2022). Los resultados determinan que la aplicación de la propuesta mejora la precisión media (mAP) en comparación con el modelo sin optimizar o la aplicación directa de la super-resolución a la imagen completa proporcionada como entrada. Modelos de detección de objetos tales como

EfficientDet D4, incrementan su precisión considerablemente, pasando de un 14.5% obtenida por el modelo pre-entrenado a un 59.1% al aplicar el enfoque optimizado. Además, los resultados experimentales muestran que aumentar el tamaño de la ventana puede conducir a mejores resultados, reduciendo así el tiempo de procesamiento necesario en algunos casos. Sin embargo, es esencial realizar un análisis preliminar basado en el área de aplicación para determinar los parámetros óptimos del enfoque propuesto. Este análisis puede garantizar que la metodología se aplique de manera óptima para lograr el mejor equilibrio entre precisión y tiempo requerido.

Finalmente, el **séptimo** trabajo se denomina *Small-Scale Urban Object Anomaly Detection Using Convolutional Neural Networks with Probability Estimation* (*Detección de Anomalías en Objetos Urbanos a Pequeña Escala Utilizando Redes Neuronales Convolucionales con Estimación de Probabilidades.* en castellano). Fue publicado en el año 2023 en la revista *SENSORS*, ocupando en dicho año la posición del segundo cuartil (100/275) de la clasificación JCR en la categoría Ingeniería, categoría de Ciencias de la Ingeniería Eléctrica y Electrónica. El artículo presenta una metodología integral que combina Redes Neuronales Convolucionales y super-resolución de imágenes para la detección de anomalías en secuencias urbanas.

Se utiliza el modelo Real-ESRGAN, diseñado para mejorar la resolución y calidad visual de imágenes individuales. El proceso se inicia procesando una secuencia en imágenes individuales, las cuales se someten a un modelo de detección de objetos, proporcionando así detecciones iniciales. Estas detecciones identifican objetos más grandes y fáciles de detectar, que se subdividen en sub-imágenes para un análisis detallado. Luego, se aplica super-resolución a estas sub-imágenes y se vuelven a pasar por el modelo de detección, generando más detecciones. Para evitar la duplicación de detecciones, se utiliza un método de agrupación basado en criterios como la superposición de cajas delimitadoras y la concordancia de clases. Las detecciones finales se seleccionan en función de las puntuaciones más altas en cada grupo. Una vez obtenidas las anotaciones, se procede a estimar la densidad de probabilidad de los elementos identificados en cada punto de la imagen. Para ello, se utiliza un estimador de densidad de probabilidad basado en núcleos. Este estimador genera una matriz que representa las zonas por las que pasan los elementos de cada clase, lo que permite detectar anomalías al establecer un umbral en la densidad de probabilidad.

Esta metodología se aplica en secuencias sintéticas creadas con el simulador CARLA Dosovitskiy et al. (2017) para detectar comportamientos anómalos en elementos de diversa clase. Los resultados obtenidos al evaluar la metodología propuesta frente a los modelos base de detección de objetos han demostrado de manera consistente mejoras notables en la identificación de elementos y un aumento significativo en las puntuaciones de *mAP*. Por ejemplo, utilizando el enfoque propuesto con el modelo *EfficientDet D4*, logró una mejora destacada en las puntuaciones de *mAP* para la clase de Peatones en la Secuencia 1, con un valor del 12.8% en comparación con el 1.8% del modelo base. Estos resultados convincentes subrayan la utilidad práctica y la robustez de la metodología propuesta en la mejora de sistemas de vigilancia y seguridad. El sistema logra una detección de anomalías más confiable y precisa al combinar de manera efectiva técnicas de super-resolución y detección de objetos. Cuando se utiliza el modelo *CenterNet HourGlass104 Keypoints 1024x1024* en la Secuencia 2, el enfoque propuesto detectó con éxito 84 anomalías, mientras que el modelo original solo identificó 4 anomalías.

B.4 Conclusiones y Trabajo Futuro

Este capítulo presenta las conclusiones de los siete trabajos que respaldan esta tesis y algunas líneas de investigación futuras motivadas por los resultados experimentales obtenidos después de probar los diferentes métodos desarrollados en ellos. El objetivo es resumir las principales contribuciones de este período de investigación y formular algunas ideas para su desarrollo futuro hacia métodos más complejos y robustos.

B.4.1 Conclusiones

El objetivo principal de esta tesis doctoral ha sido el problema de la identificación de objetos pequeños, aplicado tanto al ámbito de la detección como de la segmentación. Corresponde con un problema relativamente novedoso dentro del campo de la visión por computador. Por ello, se siguen buscando soluciones con el objetivo de mitigar el mismo. Nuestra aproximación se basa en combinar el uso de técnicas y modelos centrados en la mejora de la resolución de la imagen, para re-inferir sobre las mismas atendiendo a diversas heurísticas establecidas con el objetivo de servir como apoyo al modelo pre-entrenado de detección o segmentación de objetos independientemente, mejorando así la precisión obtenida. La mayoría de trabajos presentados se incluyen dentro de este campo. A lo largo del proceso, dado que se han adquirido los conocimientos sobre diversas problemáticas y ámbitos de forma gradual, surgieron conceptos para emplear parte de estas metodologías en otros ámbitos como era el de la detección de anomalías, representando así la parte secundaria de esta tesis doctoral.

En apoyo de esta investigación, se han incluido siete trabajos, en los cuales el autor de la tesis figura como primer autor. De estos siete trabajos, cinco se han publicado en revistas de alto impacto, dos trabajos más pequeños en los cuales se exploran ideas preliminares se han divulgado en conferencias de rango inferior.

Los trabajos que avalan esta tesis están estrechamente relacionados entre sí, dando por ello lugar al desarrollo incremental de una metodología base con el fin de mejorar la misma y ser aplicada en diversos problemas. Por ello, los trabajos que avalan esta tesis pueden ser divididos en tres grandes grupos. El primer trabajo publicado en Expert Systems (capítulo 3) corresponde con la primera aproximación de la metodología descrita para abordar el problema de la deteción de objetos pequeños. Este trabajo se centra en mejorar la precisión de diversos modelos de detección de objetos pre-entrenados mediante una metodología específica basada en la re-inferencia sobre cada una de las zonas super-resueltas generadas a partir de los elementos inicialmente detectados en secuencias donde existen deficiencias a la hora de identificar todos los elementos que aparecen en escena. La propuesta supera el rendimiento de los modelos preentrenados originales, alcanzando una precisión media del 45.1%, utilizando por ejemplo el modelo EfficientDet-D4 para la primera secuencia de vídeo. Esto representa una mejora destacada en comparación con la precisión original del modelo del 24.3%. Este enfoque se utiliza de nuevo en el trabajo publicado en la revista Pattern Recognition Letters (capítulo 6). En el primer caso, la aplicación de la metodología descrita consiguió mejorar significativamente la precisión sin necesidad de modificar el modelo. Sin embargo, el tiempo de procesamiento de las escenas se incrementaba considerablemente, pues se re-inferían sobre cada uno de los elementos iniciales detectados por la red. Por ello,

en el segundo trabajo, se aplica una etapa extra donde, atendiendo a los nuevos elementos identificados con esta metodología, el modelo de detección de objetos aprende de los mismos, consiguiendo poder identificar a elementos pequeños que inicialmente presentaban deficiencias de una sola pasada una vez se ha realizado el entrenamiento. Se puede afirmar que las métricas logradas por la propuesta presentada superan las del modelo original. Tomando la secuencia número dos como ejemplo, se observa un aumento partiendo desde un mAP promedio del 27.7% con el modelo original hasta el 92.6% después de aplicar la propuesta. Esta tendencia es consistente en las secuencias restantes. Notablemente, la aplicación de esta propuesta no solo mejora las detecciones en objetos pequeños, sino que también eleva la precisión en otros dominios.

Siguiendo esta línea, el trabajo publicado en la conferencia IWANN (capítulo 7) también se centra en la aplicación de técnicas basadas en la re-inferencia de zonas super-resueltas. Sin embargo, se modifica la metodología para detectar aquellas zonas de interés, siendo en este caso la detección de carreteras y vías de circulación a través de la binarización. Una vez identificadas las mismas, se generan áreas super-resueltas exlusivamente en dichas zonas de forma óptima a través de una heurística establecida previamente. Los resultados llevados a cabo demuestran como efectivamente, la aplicación de esta nueva metodología reduce considerablemente el número de zonas generadas y además mejora la precisión del modelo de detección de objetos. La metodología presentada produce una mejora significativa del mAP en comparación con el modelo original, con un aumento promedio de hasta el 22.7% en las secuencias de prueba evaluadas. Además, se detecta un número más sustancial de objetos, incluyendo objetos pequeños y medianos, sin necesidad de volver a entrenar el modelo de detección de objetos. Sin embargo, es posible que existan ciertos casos donde se requiera un equilibrio entre el número de áreas generadas y la precisión que logra alcanzar el modelo. De este modo, dicha línea de trabajo culmina con la publicación en Neural Computing and Applications (capítulo 8). La metodología utiliza los mismos conceptos establecidos en trabajos anteriores, sin embargo, se re-define la generación de zonas sobre las cuales aplicar super-resolución y re-inferir. Mediante el establecimiento del tamaño de ventanas deslizantes junto con el algoritmo de Bron Kerbosch para la generación máxima de cliques, se logra obtener un equilibrio entre el tiempo de procesamiento y la precisión del modelo. Entiéndase por clique a un conjunto de elementos donde cada uno se corresponde con un vehículo en secuencias de tráfico. Los resultados avalan dicha metodología, llegando a aumentar la precisión con una optimización aplicada previamente, mostrando un aumento de hasta el 44.6% en las tasas de detección. Esta transición se observa desde una tasa de detección promedio del 14.5% para el modelo EfficientDet D4 hasta un 59.1% al emplear la metodología presentada para la primera secuencia.

En paralelo, pero estrechamente relacionado, se aplican este tipo de metodologías descritas anteriormente en el ámbito de la segmentación de objetos, conformando el segundo grupo en base a los trabajos presentados. La primera de ellas fue presentada en IWINAC (capítulo 4). En ella se propone un método de re-infenrencia con super-resolución similar al presentado en el capítulo 3. Sin embargo, presenta la peculiaridad de que se trabajan con máscaras en lugar de bounding-boxes. Con ello se logra una identificación del elemento más precisa. Por ejemplo, la clase Persona aumentó de una puntuación promedia del 89.1% al 93.5%. Además, la tasa de detección de elementos mejoró del 42.71% al 71.53%. Como sucedor de los avances

desarrollados en este trabajo, surge la metodología presentada en ICAE (capítulo 5), donde de nuevo se realiza una optimización de las zonas a re-inferir y se aplica un proceso final para mejorar las máscaras de los elementos identificados múltiples veces. Los resultados con diversos modelos de segmentación de objetos avalan la mejora en la percisión tras aplicar la metodología descrita, revelando mejoras de hasta el 8.1% cuando se implementa con el modelo YOLACT++ en una de las secuencias.

Finalmente, el tercer grupo de trabajos se centra en la detección de anomalías. El trabajo presentado en *SENSORS* (capítulo 9) utiliza una metodología similar a la descrita en trabajos como el presentado en el capítulo 7 para mejorar la identificación de características. Sin embargo, para identificar posteriormente los elementos situados en regiones inusuales, se calculan las regiones comunes de cada clase. Por ejemplo, haciendo uso del enfoque presentado y utilizando el modelo EfficientDet D4, se demostró una mejora notable en los mAP para la clase de peatones en la Secuencia 1, alcanzando un valor del 12.8% en comparación con el obtenido por el modelo base, el cual fue de un 1.8%.

Estos trabajos han abordado el problema de forma diferente utilizando herramientas y paradigmas similares, los cuales permiten resolver el mismo, mejorando por ello principalmente la detección de objetos pequeõs.

B.5 Trabajo Futuro

De acuerdo con la investigación realizada en la elaboración de esta tesis doctoral, se extraen una serie de conclusiones y líneas de investigación futuras con el objetivo de continuar con los desarrollos llevados a cabo. Por ello, se determinan las vías por las cuales mejorar la propuesta presentada, así como la formulación de nuevas ideas para ser aplicadas en nuevos enfoques. Como conclusión de dicha tesis doctoral, es necesario destacar que el problema de la detección y segmentación de objetos pequeños, pese a que el campo del aprendizaje profundo ha supuesto una revolución en multitud de ámbitos aportando por ello soluciones ante problemas complejos, existen problemas pendientes de mejora pues tanto la detección como segmentación de este tipo de elementos es deficiente en multitud de casos, constituyendo una tarea compleja. En esta tesis se realizan una serie de trabajos que intentan mitigar dicho problema, obteniendo por ello buenos resultados. Es por ello por lo que puede ser particularmente atractivo seguir trabajando en esta dirección con el fin de aportar soluciones más robustas.

De acuerdo con los trabajos que avalan esta tesis, las líneas futuras se dividen en tres grandes grupos atendiendo a la temática planteada:

- Continuar trabajando en los modelos de detección de objetos para mejorar más si cabe la precisión de los modelos.
- Proseguir en el ámbito de la segmentación aportando nuevas técnicas que mejoren la precisión de las máscaras obtenidas por el modelo.
- Ampliar el estudio de detección de anomalías para abarcar nuevos ámbitos como por ejemplo las anomalías de trayectoria en secuencias de vídeo.

B.5.1 Mejora en la detección de objetos pequeños

La primera línea de trabajo más viable a corto plazo se basa en el desarrollo y mejora de la metodología desarrollada en los trabajos propuestos a lo largo de la tesis para mejorar la tasa de precisión. Por ello, existen tres puntos donde se establecen las posibles mejoras a realizar:

- Detección de objetos en conjunto: Se podría experimentar con una amplia gama de modelos centrados en la detección de objetos con diversas arquitecturas con el objetivo de combinar las salidas aportadas por cada uno de los modelos y generar una solución unificada aportando robustez.
- Selección de características: Investigar y aplicar otros algoritmos de selección de características con el fin de mejorar la precisión de la localización de objetos en escenas estáticas. Gracias a este tipo de algoritmos, sería posible identificar características más descriptivas para la posterior detección de los objetos en dichas regiones.
- Integración de la información temporal: Mediante el desarrollo de técnicas para la integración de la información temporal en la detección de objetos en secuencias de vídeo estáticas. Alguna de estas técnicas corresponderían con filtros Kalman.
- Aplicar Otras Técnicas para Mejorar la Calidad de la Imagen: Aplicar técnicas avanzadas de mejora de imagen, como la reducción de ruido, el realce de contraste y la corrección de color, podría mejorar la calidad de la imagen más allá de la super-resolución, mejorando así la precisión en la detección de objetos y el análisis de escenas.

B.5.2 Mejora en la segmentación de objetos pequeños

La segunda línea basada en la mejora en la segmentación de objetos pequeños constituye de nuevo una de las vías más viables para proseguir con el desarrollo de nuevas técnicas que puedan mejorar la metodología propuesta. Para ello, se establecen las siguientes posibles mejoras:

- Técnicas Avanzadas de Selección de Características: Explorar e implementar a la metodología propuesta métodos avanzados basados en la selección de características más allá de SIFT o SURF en entre otros con el fin de priorizar características relevantes en la escena para determinar así las áreas sobre las cuales incidir cone el fin de obtener una segmentación más precisa.
- Estrategias de Aumento de Datos y Aprendizaje por Transferencia Avanzado: Aplicar diversas técnicas de aumento de datos adaptadas a tareas de segmentación, tales como una serie de transformaciones que simulen determinadas condiciones o la generación de imágenes sintéticas mediante el uso de redes generativas GAN de acuerdo a los datos de entrada podría dar lugar a un mejor entrenamiento. Realizar *Transfer-Learning* con modelos previamente entrenados para mejorar la segmentación de objetos. Se podría realizar un análisis en base al ajuste de los parámetros del modelo para así determinar cuál de ellos funciona mejor en una serie de dominios.

B.5.3 Detección de anomalías

Aunque el ámbito de la tesis se ha centrado en la mejora en la detección y segmentación de objetos pequeños, también se han aplicado partes de esta metodología para la identificación de anomalías. Por ello, las vías planteadas son las siguientes:

- Mejorar la Robustez en Condiciones de Iluminación Desafiantes: Investigar y desarrollar aún más técnicas para mejorar el rendimiento de la metodología en escenarios con condiciones de iluminación adversas. Esto podría implicar el desarrollo de métodos avanzados de corrección de iluminación más allá de las Tablas de Búsqueda (LUTs), como enfoques basados en aprendizaje profundo o control de exposición adaptativo.
- Pruebas y Validación en el Mundo Real: Realizar pruebas y validación exhaustivas en diversos entornos de vigilancia del mundo real, incluyendo entornos urbanos, rurales y remotos. Recopilar datos bajo diferentes condiciones de iluminación y escenarios para evaluar la confiabilidad y efectividad de la metodología.

Bibliography

Nobody knows what's gonna happen at the end of the line, so you might as well enjoy the trip.

Grim Fandango, Manuel Calavera.

Ahmed, M., Hashmi, K. A., Pagani, A., Liwicki, M., Stricker, D. and Afzal, M. Z. Survey and performance analysis of deep learning based object detection in challenging environments. *Sensors*, vol. 21(15), 2021. ISSN 1424-8220.

Akyon, F. C., Onur Altinuc, S. and Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970. 2022.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Esesn, B. C. V., Awwal, A. A. S. and Asari, V. K. The history began from alexnet: A comprehensive survey on deep learning approaches. 2018.

B. Sophia, D. C. Segmentation based real time anomaly detection and tracking model for pedestrian walkways. *Intelligent Automation & Soft Computing*, vol. 36(3), pages 2491–2504, 2023. ISSN 2326-005X.

Bai, Y., Zhang, Y., Ding, M. and Ghanem, B. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Computer Vision – ECCV 2018* (edited by V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss), pages 210–226. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01261-8.

Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, vol. 110(3), pages 346–359, 2008. ISSN 1077-3142. Similarity Matching in Computer Vision and Multimedia.

Bay, H., Tuytelaars, T. and Van Gool, L. Surf: Speeded up robust features. In *Computer Vision – ECCV 2006* (edited by A. Leonardis, H. Bischof and A. Pinz), pages 404–417. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-33833-8.

BOLYA, D., ZHOU, C., XIAO, F. and LEE, Y. J. Yolact: Real-time instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9156–9165. 2019.

BOLYA, D., ZHOU, C., XIAO, F. and LEE, Y. J. Yolact++ better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44(2), pages 1108–1121, 2022.

CHANG, C.-W., CHANG, C.-Y. and LIN, Y.-Y. A hybrid cnn and lstm-based deep learning model for abnormal behavior detection. *Multimedia Tools and Applications*, vol. 81(9), pages 11825–11843, 2022. ISSN 1573-7721.

CHEN, C., LIU, M.-Y., TUZEL, O. and XIAO, J. R-cnn for small object detection. In *Computer Vision – ACCV 2016* (edited by S.-H. Lai, V. Lepetit, K. Nishino and Y. Sato), pages 214–230. Springer International Publishing, Cham, 2017. ISBN 978-3-319-54193-8.

CONG, Y., YUAN, J. and LIU, J. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. 2011.

CORDTS, M., OMRAN, M., RAMOS, S., REHFELD, T., ENZWEILER, M., BENENSON, R., FRANKE, U., ROTH, S. and SCHIELE, B. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223. 2016.

DALAL, N. and TRIGGS, B. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pages 886–893 vol. 1. 2005.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. and FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. 2009.

DONG, C., LOY, C. C., HE, K. and TANG, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38(2), pages 295–307, 2016a. ISSN 0162-8828.

DONG, C., LOY, C. C. and TANG, X. Accelerating the super-resolution convolutional neural network. In *Computer Vision – ECCV 2016* (edited by B. Leibe, J. Matas, N. Sebe and M. Welling), pages 391–407. Springer International Publishing, Cham, 2016b. ISBN 978-3-319-46475-6.

DOSOVITSKIY, A., ROS, G., CODEVILLA, F., LOPEZ, A. and KOLTUN, V. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning* (edited by S. Levine, V. Vanhoucke and K. Goldberg), vol. 78 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2017.

DU, D., QI, Y., YU, H., YANG, Y., DUAN, K., LI, G., ZHANG, W., HUANG, Q. and TIAN, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Computer Vision – ECCV 2018* (edited by V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss), pages 375–391. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01249-6.

DUAN, K., BAI, S., XIE, L., QI, H., HUANG, Q. and TIAN, Q. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6568–6577. 2019.

DURAISAMY, P., NARAYANAN, V. B. S., PATTURAJAN, R. and VEERASAMY, K. *Multi-sensor Fusion Methods for Unmanned Aerial Vehicles to Detect Environment Using Deep Learning Techniques*, pages 263–273. Springer International Publishing, Cham, 2022. ISBN 978-3-030-97113-7.

ELHARROUSS, O., ALMAADEED, N. and AL-MAADEED, S. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, vol. 77, page 103116, 2021. ISSN 1047-3203.

GARCÍA-AGUILAR, I., DEKA, L., LUQUE-BAENA, R. M., DOMÍNGUEZ, E. and LÓPEZ-RUBIO, E. Minimal optimal region generation for enhanced object detection in aerial images using super-resolution and convolutional neural networks. In *Advances in Computational Intelligence*, pages 276–287. Springer Nature Switzerland, 2023.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M. and LÓPEZ-RUBIO, E. Automated labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks. *Pattern Recognition Letters*, vol. 167, pages 45–52, 2023. ISSN 0167-8655.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M. and LÓPEZ-RUBIO, E. Object detection in traffic videos: an optimized approach using super-resolution and maximal clique algorithm. *Neural Computing and Applications*, vol. 35(26), pages 18999–19013, 2023a. ISSN 1433-3058.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M., LÓPEZ-RUBIO, E. and DOMÍNGUEZ, E. Optimized instance segmentation by super-resolution and maximal clique generation. *Integrated Computer-Aided Engineering*, vol. 30, pages 243–256, 2023b. ISSN 1875-8835. 3.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M., LÓPEZ-RUBIO, E. and DOMÍNGUEZ-MERINO, E. Enhanced image segmentation by a novel test time augmentation and super-resolution. In *Bio-inspired Systems and Applications: from Robotics to Ambient Intelligence* (edited by J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López and H. Adeli), pages 153–162. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06527-9.

GARCÍA-AGUILAR, I., LUQUE-BAENA, R. M., DOMÍNGUEZ, E. and LÓPEZ-RUBIO, E. Small-scale urban object anomaly detection using convolutional neural networks with probability estimation. *Sensors*, vol. 23(16), 2023. ISSN 1424-8220.

GARCÍA-AGUILAR, I., LUQUE-BAENA, R. M. and LÓPEZ-RUBIO, E. Improved detection of small objects in road network sequences using cnn and super resolution. *Expert Systems*, vol. 39(2), page e12930, 2022.

GAYAL, B. S. and PATIL, S. R. Detection and localization of anomalies in video surveillance using novel optimization based deep convolutional neural network. *Multimedia Tools and Applications*, vol. 82(19), pages 28895–28915, 2023. ISSN 1573-7721.

GHIASI, G., CUI, Y., SRINIVAS, A., QIAN, R., LIN, T.-Y., CUBUK, E. D., LE, Q. V. and ZOPH, B. Simple copy-paste is a strong data augmentation method for instance segmentation. pages 2917–2927. IEEE, 2021. ISBN 978-1-6654-4509-2.

GIRSHICK, R. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448. IEEE Computer Society, Los Alamitos, CA, USA, 2015. ISSN 2380-7504.

GIRSHICK, R., DONAHUE, J., DARRELL, T. and MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587. 2014.

GUERRERO-GÓMEZ-OLMEDO, R., LÓPEZ-SASTRE, R. J., MALDONADO-BASCÓN, S. and FERNÁNDEZ-CABALLERO, A. Vehicle tracking by simultaneous detection and viewpoint estimation. In *Natural and Artificial Computation in Engineering and Medical Applications* (edited by J. M. Ferrández Vicente, J. R. Álvarez Sánchez, F. de la Paz López and F. J. Toledo Moreo), pages 306–316. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-38622-0.

HAYDER, Z., HE, X. and SALZMANN, M. Boundary-aware instance segmentation. pages 587–595. IEEE, 2017. ISBN 978-1-5386-0457-1.

HE, K., GKIOXARI, G., DOLLÁR, P. and GIRSHICK, R. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. 2017a.

HE, K., GKIOXARI, G., DOLLAR, P. and GIRSHICK, R. Mask r-cnn. pages 2980–2988. IEEE, 2017b. ISBN 978-1-5386-1032-9.

HU, X., DAI, J., HUANG, Y., YANG, H., ZHANG, L., CHEN, W., YANG, G. and ZHANG, D. A weakly supervised framework for abnormal behavior detection and localization in crowded scenes. *Neurocomputing*, vol. 383, pages 270–281, 2020. ISSN 0925-2312.

IOFFE, S. and SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

JOCHER, G., CHAURASIA, A. and QIU, J. YOLO by Ultralytics. 2023.

KANU-ASIEGBU, A. M., VASUDEVAN, R. and DU, X. Leveraging trajectory prediction for pedestrian video anomaly detection. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–08. 2021.

KEMBHAVI, A., HARWOOD, D. and DAVIS, L. S. Vehicle detection using partial least squares. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33(6), pages 1250–1265, 2011.

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J. and Cho, K. Augmentation for small object detection. 2019.

Krishna, H. and Jawahar, C. Improving small object detection. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 340–345. 2017.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, vol. 60, pages 84–90, 2017. ISSN 0001-0782.

Lai, W.-S., Huang, J.-B., Ahuja, N. and Yang, M.-H. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843. 2017.

Laradji, I. H., Rostamzadeh, N., Pinheiro, P. O., Vazquez, D. and Schmidt, M. Proposal-based instance segmentation with point supervision. pages 2126–2130. IEEE, 2020. ISBN 978-1-7281-6395-6.

Laradji, I. H., Vázquez, D. and Schmidt, M. W. Where are the masks: Instance segmentation with image-level supervision. *BMVC*, vol. abs/1907.01430, 2019.

Le, T. T., Tran, S. T., Mita, S. and Nguyen, T. D. Real time traffic sign detection using color and shape-based features. In *Intelligent Information and Database Systems* (edited by N. T. Nguyen, M. T. Le and J. Świątek), pages 268–278. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12101-2.

Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86(11), pages 2278–2324, 1998.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114. 2017.

Lim, B., Son, S., Kim, H., Nah, S. and Lee, K. M. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140. 2017.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944. 2017.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C. Ssd: Single shot multibox detector. In *Computer Vision – ECCV 2016* (edited by B. Leibe, J. Matas, N. Sebe and M. Welling), pages 21–37. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46448-0.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60(2), pages 91–110, 2004. ISSN 1573-1405.

LYU, S., CHANG, M.-C., DU, D., LI, W., WEI, Y., COCO, M. D., CARCAGNÌ, P., SCHUMANN, A., MUNJAL, B., DANG, D.-Q.-T., CHOI, D.-H., BOCHINSKI, E., GALASSO, F., BUNYAK, F., SEETHARAMAN, G., BAEK, J.-W., LEE, J. T., PALANIAPPAN, K., LIM, K.-T., MOON, K., KIM, K.-J., SOMMER, L., BRANDLMAIER, M., KANG, M.-S., JEON, M., AL-SHAKARJI, N. M., ACATAY, O., KIM, P.-K., AMIN, S., SIKORA, T., DINH, T., SENST, T., CHE, V.-G.-H., LIM, Y.-C., SONG, Y.-M. and CHUNG, Y.-S. Ua-detrac 2018: Report of avss2018 & iwt4s challenge on advanced traffic monitoring. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. 2018.

LYU, S., CHANG, M.-C., DU, D., WEN, L., QI, H., LI, Y., WEI, Y., KE, L., HU, T., DEL COCO, M., CARCAGNÌ, P., ANISIMOV, D., BOCHINSKI, E., GALASSO, F., BUNYAK, F., HAN, G., YE, H., WANG, H., PALANIAPPAN, K., OZCAN, K., WANG, L., WANG, L., LAUER, M., WATCHARAPINCHAI, N., SONG, N., AL-SHAKARJI, N. M., WANG, S., AMIN, S., RUJIKIETGUMJORN, S., KHANOVA, T., SIKORA, T., KUTSCHBACH, T., EISELEIN, V., TIAN, W., XUE, X., YU, X., LU, Y., ZHENG, Y., HUANG, Y. and ZHANG, Y. Ua-detrac 2017: Report of avss2017 & iwt4s challenge on advanced traffic monitoring. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. 2017.

MAHADEVAN, V., LI, W., BHALODIA, V. and VASCONCELOS, N. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. 2010.

MOUMENE, I. and OUELAA, N. Gears and bearings combined faults detection using optimized wavelet packet transform and pattern recognition neural networks. *The International Journal of Advanced Manufacturing Technology*, vol. 120(7), pages 4335–4354, 2022. ISSN 1433-3015.

PUSTOKHINA, I. V., PUSTOKHIN, D. A., VAIYAPURI, T., GUPTA, D., KUMAR, S. and SHANKAR, K. An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety. *Safety Science*, vol. 142, page 105356, 2021. ISSN 0925-7535.

REDMON, J., DIVVALA, S., GIRSHICK, R. and FARHADI, A. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788. 2016.

REDMON, J. and FARHADI, A. Yolov3: An incremental improvement. 2018.

REN, S., HE, K., GIRSHICK, R. and SUN, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39(6), pages 1137–1149, 2017.

RONNEBERGER, O., FISCHER, P. and BROX, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (edited by N. Navab, J. Hornegger, W. M. Wells and A. F. Frangi), pages 234–241. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24574-4.

Sabokrou, M., Fathy, M., Hoseini, M. and Klette, R. Real-time anomaly detection and localization in crowded scenes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 56–62. 2015.

Shen, M., Jiang, X. and Sun, T. Anomaly detection by analyzing the pedestrian behavior and the dynamic changes of behavior. In *Intelligent Computing Theories and Application* (edited by D.-S. Huang, V. Bevilacqua, P. Premaratne and P. Gupta), pages 211–222. Springer International Publishing, Cham, 2017. ISBN 978-3-319-63309-1.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D. and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883. 2016.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, vol. 15(1), pages 1929–1958, 2014. ISSN 1532-4435.

Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pages 6105–6114. 2019.

Tan, M., Pang, R. and Le, Q. V. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787. 2020.

Tseng, C.-H., Hsieh, C.-C., Jwo, D.-J., Wu, J.-H., Sheu, R.-K. and Chen, L.-C. Person retrieval in video surveillance using deep learning–based instance segmentation. *Journal of Sensors*, vol. 2021, page 9566628, 2021. ISSN 1687-725X.

U.S. Department Of Transportation Federal Highway Administration. Next generation simulation (ngsim) vehicle trajectories and supporting data. 2017. `https://data.transportation.gov/d/8ect-6jqj`.

Wang, X., Xie, L., Dong, C. and Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914. 2021.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. and Loy, C. C. Esrgan: Enhanced super-resolution generative adversarial networks. In *Computer Vision – ECCV 2018 Workshops* (edited by L. Leal-Taixé and S. Roth), pages 63–79. Springer International Publishing, Cham, 2019. ISBN 978-3-030-11021-5.

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.-C., Qi, H., Lim, J., Yang, M.-H. and Lyu, S. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, vol. 193, page 102907, 2020. ISSN 1077-3142.

WU, Y., KIRILLOV, A., MASSA, F., LO, W.-Y. and GIRSHICK, R. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

YANG, Q., DONG, E. and ZHU, L. An instance segmentation algorithm based on improved mask r-cnn. In *2020 Chinese Automation Congress (CAC)*, pages 4804–4809. 2020.

ZHANG, Y., CHU, J., LENG, L. and MIAO, J. Mask-refined r-cnn: A network for refining object details in instance segmentation. *Sensors*, vol. 20(4), 2020. ISSN 1424-8220.

ZHAO, B., FEI-FEI, L. and XING, E. P. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. 2011.

ZHU, P., WEN, L., DU, D., BIAN, X., FAN, H., HU, Q. and LING, H. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44(11), pages 7380–7399, 2022.

ZHU, X., LYU, S., WANG, X. and ZHAO, Q. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2778–2788. 2021.

No Matter What, You Keep Finding Something To Fight For.

The Last of Us, Joel Miller.