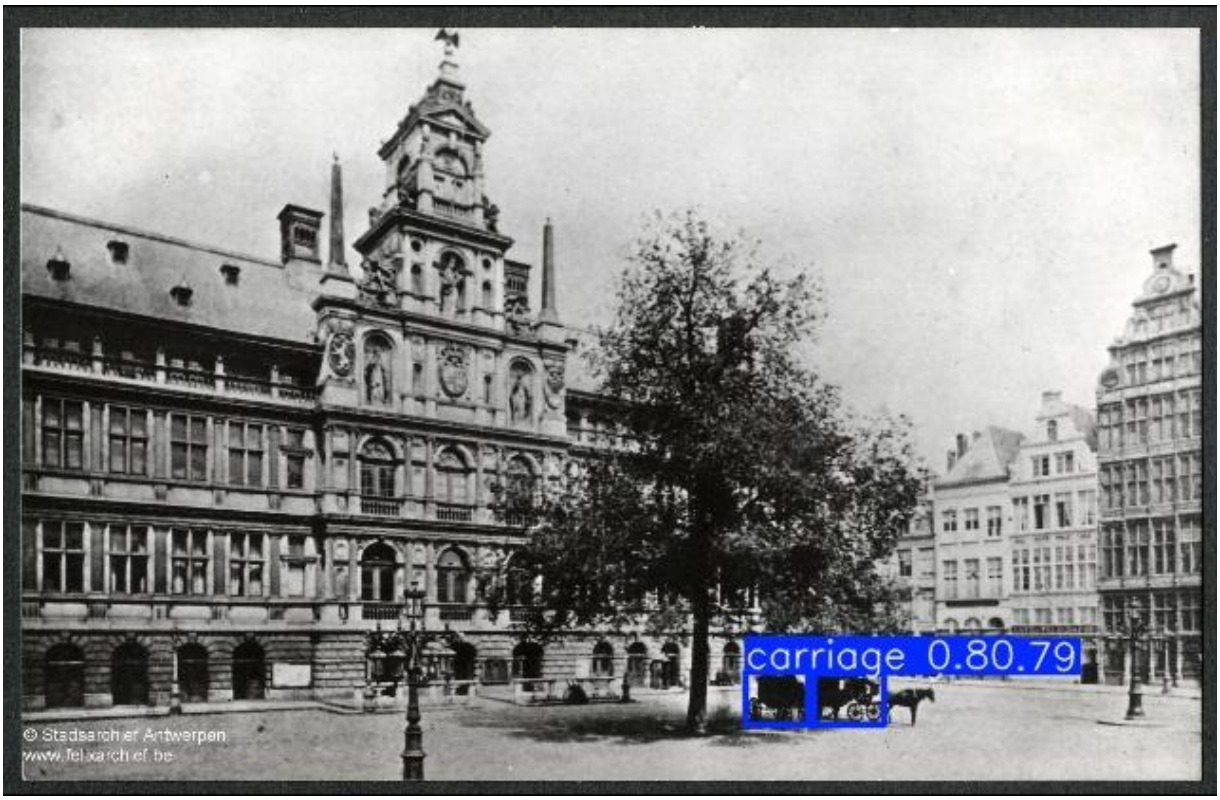
Universiteit Antwerpen
Faculteit Ontwerpwetenschappen

The Sound of the City

Object Recognition of Sound Sources on Historical Photographs



Cuykens, Vastert

20192567

Dr. Hasan Baran Firat

Prof. Piraye Hacigüzeller

Universiteit Antwerpen
Faculteit Ontwerpwetenschappen
Master in de Erfgoedstudies
Academiejaar 2024-2025

UA_2051FOWERF-2425
6 juni 2025
Aantal woorden: 21348

Abstract English

This thesis explores how machine learning can be used to detect sound sources in historical photographs, with the aim of developing a model that can be used for historical soundscape reconstruction. While machine learning, and more specifically object detection, is increasingly being used in heritage studies, its application on historical photographs is largely unexplored.

The research is divided into four stages. First, three object detection models, YOLOv8, Faster R-CNN, and RetinaNet, were benchmarked on both modern and historical datasets to assess their baseline performance, without any training. Although all models performed significantly worse on historical photographs, YOLOv8 showed the smallest drop in accuracy. In the second and third parts, YOLOv8 was fine-tuned on two classes of sound-producing objects: an existing COCO class, “train”, and a newly added class, “carriage”. In total, over 1600 photographs were annotated for the various datasets. While the modest dataset size and limited computational resources influenced the performance of the custom models, the research nevertheless offers some interesting insights into the potential and limitations of applying object detection to historical photographs. In the last stage, a custom CycleGAN model was trained to transform modern images into a historical style and vice versa, helping to generate training data and reduce the need for manual annotation. The thesis demonstrates the feasibility of using object detection on historical photographs for historical soundscape research and highlights promising directions for future work on machine learning and heritage studies.

Abstract Nederlands

In deze scriptie wordt onderzocht hoe *machine learning* gebruikt kan worden om geluidsbronnen in historische foto’s te detecteren, met als doel een model te ontwikkelen dat gebruikt kan worden voor de reconstructie van historische *soundscapes*. Hoewel *machine learning*, en meer specifiek *object detection*, steeds vaker worden gebruikt binnen erfgoedstudies, is de toepassing ervan op historische foto’s nog grotendeels onontgonnen terrein.

Het onderzoek is onderverdeeld in vier stukken. Eerst werden drie *object detection models*, namelijk YOLOv8, Faster R-CNN en RetinaNet, gebenchmarkt op zowel moderne als historische datasets om zo hun basisprestaties te beoordelen, zonder enige training. Hoewel alle modellen aanzienlijk slechter presteerden op historische foto’s was de daling in nauwkeurigheid bij YOLOv8 het kleinst. In het tweede en derde deel werd YOLOv8 *gefinetuned* op twee soorten objecten, namelijk een bestaande COCO-klasse, “trein”, en een nieuw toegevoegde klasse, “koets”. In totaal werden meer dan 1600 foto’s geannoteerd voor de verschillende datasets. Hoewel de eerder bescheiden omvang van de datasets en de bepekte middelen de prestaties van de modellen hebben beïnvloed, biedt het onderzoek niettemin enkele interessante inzichten in de mogelijkheden en beperkingen van het toepassen van *object detection* op

historische foto's. In de laatste fase werd een CycleGAN-model getraind om hedendaagse foto's om te vormen naar een historische stijl en omgekeerd, met als doel de arbeidsintensieve aard van manuele annotatie te verlichten. Deze thesis toont de haalbaarheid aan van het gebruik van *object detection* op historische foto's en wijst op enkele veelbelovende mogelijkheden voor toekomstig onderzoek binnen *machine learning* en erfgoedstudies.

Content

List of tables

List of figures

Glossary

Term	Definition
Annotation	The process of labelling objects in an image with bounding boxes and class names.
Augmentation Technique	A technique to artificially enlarge datasets by applying various transformations to images.
Average Precision	A metric used to evaluate the performance of object detection models. It gives a balanced assessment of precision and recall by considering the area under the precision-recall curve. The closer the value to 1, the better the model.
Batch Size	The number of images used in one forward and backward pass through the network.
Bounding Box	A box drawn around the target object in an image to mark its location.
Microsoft Common Objects in Context (COCO)	A widely used dataset containing labelled images of different objects in natural settings. It contains images of 80 different classes.
Convolutional Neural Networks (CNNs)	A class of deep neural networks mainly designed for tasks that use object recognition, image classification and segmentation.
CycleGAN	A type of Generative Adversarial Network (GAN) that performs unpaired image-to-image translation, enabling style transfer of images between domains.
Discriminator	A part of the architecture of a GAN. It takes the input data and the new data produced by the Generator and attempts to distinguish between

	the two and outputs the probability that the data is real.
Epoch	A complete pass through the entire dataset.
False Negative (FN)	An object that was present in the picture but not detected by the model.
False Positive (FP)	When the model predicts an object that is not present in the picture or when the bounding box does not accurately correspond with the ground truth.
Faster R-CNN	A two-stage object detection model characterised by the use of a Region Proposal Network (RPN) and Convolutional Neural Networks (CNNs). It first proposes regions of interest and then classifies and refines the bounding boxes.
Fine-tuning	A technique of transfer learning where a pretrained model is further trained on a new dataset to adapt it to a specific task.
Generator	A part of the architecture of a GAN. The goal of the Generator is to produce new data that resembles the input data as close as possible
Graphics Processing Unit (GPU)	An electronic circuit designed to speed up the creation of images and videos. It is used in machine learning because it excels at handling computational demanding tasks.
Hyperparameters	Configuration settings (like learning rate, batch size, number of epochs) that influence the model's training and performance.

Intersection over Union (IoU)	The ratio of the intersection area to the union area of the predicted bounding box and the ground truth bounding box. When the IoU is 0, the two boxes do not overlap, when it is 1, the two boxes are identical.
Loss Function	A metric that measures the model performance by calculating the difference between a predicted value and the ground truth.
Mean Average Precision	A metric used to evaluate the performance of object detection models. It averages the average precision values across all the classes or across multiple intersection over union thresholds. The closer the value to 1, the better the model.
Object Detection	A task within computer vision that involves identifying and localising objects within an image by drawing bounding boxes around it and classifying the object.
Overfitting	An issue where the model adapts too closely to the training data. It achieves excellent performance on the training data but performs significantly worse on unseen data.
Precision	A metric used to evaluate how precise the model is in its predictions. It measures the proportion of correctly predicted objects (true positives) out of all predicted objects (true positives and false positives). $$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Preprocessing	The set of transformations applied to raw data to prepare it for model training, like resizing and normalisation.
Recall	A metric used to evaluate the ability of the model to find all relevant objects present in the image. It is the percentage of correct positive predictions among all given ground truths. $$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$
RetinaNet	A single-stage object detection model known for using a Feature Pyramid Network (FPN) for multi-scale feature extraction and a Focal Loss function to handle class imbalance.
Single-stage Detector	A class of object detection models that directly predicts bounding boxes and class probabilities in one step.
Tensor Processing Unit (TPU)	An application-specific integrated circuit custom-developed by Google used to accelerate machine learning workload.
Transfer Learning	A technique where an existing object detection model trained on large datasets for a specific task is reused and trained with a smaller dataset for a different task.
True Positive (TP)	An object correctly detected and classified by the model.
Two-stage detector	A class of object detection models that first generates region proposals and then classifies and refines these proposals.

Underfitting	An issue where the model does not perform well on both the training datasets as on the unseen data.
YOLO	A family of real-time, single-stage object detection models that revolutionised the field by introducing a model that used a grid-based approach to predict bounding boxes and class probabilities at the same time.

1. Introduction

"We have no ear lids. We are condemned to listen. But this does not mean our ears are always open."[1] With this quote, Canadian composer, writer and acoustic ecologist R. Murray Schafer sought to draw attention to the often passive nature of our relationship with sound. We are constantly surrounded by sounds, but we do not always actively listen to them. This quote, published in 2003, reflects Schafer's lifelong exploration of the auditory environment. While the term "soundscape" was introduced before Schafer, he is widely credited with popularising and theorising it in a way that laid the foundations for a new field.[2] Since his influential publication from 1977, *The Soundscape: our sonic environment and the tuning of the world*, disciplines ranging from urban planning to architecture started to devote attention to the auditory aspects of everyday life. [3]

In the field of heritage studies, the term soundscapes has only recently been incorporated into research. Historically, there has been a dominance of vision-based approaches in heritage studies, leaving other senses in the background.[4] Hasan Baran Firat argued that the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage and the introduction of the term Intangible Cultural Heritage (ICH) opened the door for challenging this "vision-centric" focus.[5] However, despite the attempt to broaden the sensory scope of heritage, the bias toward the visual sense remained persistent. ICH opened up space for non-visual senses, but it did not fully conceive unique sensory heritage attitudes for each modality.[6]

Despite the lagging behind of the inclusion of other senses in heritage, it can be observed that more and more attention is nevertheless being given to the non-visual. Dan Luo et al. state that the introduction of the idea of *sensory museology* by David Howes in 2014 has resulted in research in exhibition design and

---

[1] R. Murray Schafer, "Open Ears," *Soundscape: The Journal of Acoustic Ecology* 4, no. 2 (2003): 14.
[2] Hasan Baran Firat and Luigi Maffei, "A methodology for the historically informed soundscape" (Inter-.Noise 2020, e-congress, 2020).
[3] R. Murray Schafer, *The Soundscape: our sonic environment and the tuning of the world* (New York: Alfred Knopf, Inc., 1977); Jian Kang et al., "Ten Questions on the Soundscapes of the Built Environment," *Building and Environment* 108 (2016): 285, https://doi.org/10.1016/j.buildenv.2016.08.011; Firat and Maffei, "Short A methodology for the historically informed soundscape." 3.
[4] Hasan Baran Firat, "Acoustics as Tangible Heritage: Re-embodying the Sensory Heritage in the Boundless Reign of Sight," *Preservation, Digital Technology & Culture* 50, no. 1 (2021): 4-6, https://doi.org/10.1515/pdtc-2020-0028; Johannes Müske, "Constructing Sonic Heritage: The Accumulation of Knowledge in the Context of Sound Archives," *The Journal of Ethnology and Folkloristics* 4, no. 1 (2011): 39.
[5] Firat, "Acoustics as Tangible Heritage: Re-embodying the Sensory Heritage in the Boundless Reign of Sight," 4; UNESCO, *Basic Texts of the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage*, 2022 ed. (Paris: France, 2022).
[6] Firat, "Acoustics as Tangible Heritage: Re-embodying the Sensory Heritage in the Boundless Reign of Sight," 4.

in understanding factors influencing visitor responses, from sensory stimuli to embodied cognition.[7] Large-scale projects like the European research project Odeuropa, which explores the role of scents in European cultural heritage, have also helped to put the spotlight on sensory heritage.[8] Research has demonstrated strong connections between emotions, sensory experiences, and long-term memory, making this multi-sensory approach a valuable perspective in heritage studies.[9]

With regards to hearing, UNESCO recognised the importance of sounds in society in 2017, illustrated by the publication of resolution 39C/49, titled *The Importance of Sound in Today's World: Promoting Best Practices*.[10] With this resolution, UNESCO not only addressed different issues regarding sound, like noise pollution, but also noted the importance of sound recording, reproduction and conservation technology.[11] They stated that: "the sound environment reflects and shapes our individual and collective behaviour, and our productivity and capacity to live in harmony together."[12] This importance of soundscapes for societies is a recurring theme in research. The combination of different sounds can create culturally distinct soundscapes, that are representative of specific communities.[13] As Pinar Yelmi stated: "Sounds are powerful values that remind people where they come from, their origins and memories."[14] It is not surprising then that different projects emerged that want to protect soundscapes linked to specific communities or monuments from disappearance by recording and archiving them.

In 2015 for example, Stuart Fowkes, a sound artist and field recordist, founded the sound project "Cities and Memory", where people from all over the world can upload field recordings from specific places, creating a large sound archive and sound map.[15] Currently, there are more than 7000 sounds collected, spread over 130 countries and territories.[16] For World Heritage Day 2025, the project devoted extra attention to the sounds from UNESCO World Heritage sites and sounds from the list of Intangible Cultural

---

[7] Dan Luo, Lieve Doucé, and Karin Nys, "Multisensory museum experience: an integrative view and future research directions," *Museum Management and Curatorship* (2024), https://doi.org/10.1080/09647775.2024.2357071; David Howes, "Introduction to Sensory Museology," *The Senses and Society* 9, no. 3 (2014), https://doi.org/10.2752/174589314X14023847039917.
[8] "Smell Heritage – Sensory Mining," accessed 30 March 2025, https://odeuropa.eu/.
[9] Francesco Galvano, "The Triad of Senses, Emotions, and Memory: Dynamic Interactions and Multidisciplinary Implications," (2015); Ana Bender et al., "Sensory experiences in heritage contexts: A qualitative approach.," *European Journal of Tourism Research* 36 (2024): 2, https://doi.org/10.54055/ejtr.v36i.3060.
[10] UNESCO, "39 C/49 The Importance of Sound in Today's World: Promoting Best Practices" (General Conference, 39th session, Paris 2017).
[11] Ibid.
[12] Ibid.
[13] Eva Pietroni, "Mapping the Soundscape in Communicative Forms for Cultural Heritage: Between Realism and Symbolism," *Heritage* 4, no. 4 (2021): 4497, https://doi.org/10.3390/heritage4040248.
[14] Pinar Yelmi, "Protecting Contemporary Cultural Soundscapes as Intangible Cultural Heritage: Sounds of Istanbul," *International Journal of Heritage Studies* 22, no. 4 (2016): 309, https://doi.org/10.1080/13527258.2016.1138237.
[15] "What is Cities and Memory?," Cities and Memory, Oxford, accessed 31 March 2025, https://citiesandmemory.com/what-is-cities-and-memory-about/.
[16] Ibid.

Heritage. The initiators of the project stated that sounds are now overlooked in tourism, even though sounds offer an alternative perspective on how we experience cultural sites.[17] Despite their significance, sounds thus remain largely absent from heritage conservation efforts, something that they want to battle with this project.[18]

These crowdsource initiatives are a popular method for collecting important sounds. In her doctoral research, Pinar Yelmi established two archives based on crowdsource methods.[19] Yelmi wanted to protect sounds from the city of Istanbul, where, because of its dynamic nature, daily life and urban sounds change rapidly.[20] Based on interviews and online surveys, sound symbols of the city were determined and recorded with professional equipment. These sounds were then collected under the name "The Soundscape of Istanbul" and were made publicly accessible at Koç University Suna Kıraç Library and on Europeana.[21] To further expand the collection, a second crowd-based project was launched. With this "Soundsslike" project, an interactive online archive was made where anyone could upload own sound recordings.[22]

But what about soundscapes that no longer exist? Schafer himself noted the difficulty of studying those sounds that are lost.[23] In his words: "We are also disadvantaged in the pursuit of a historical perspective […] sounds may alter or disappear with scarcely a comment even from the most sensitive of historians."[24] Schafer argues that different types of sources need to be used if we want to take a look at historical soundscapes. In his opinion, "earwitness" accounts from literature and mythology, as well as anthropological and historical records are the go-to sources for historical soundscapes.[25] While Schafer mainly focuses on written sources, Jane Malcolm-Davies argued in her article about historical

[17] "Sonic Heritage - Exploring the sounds of the world's most famous sights," Cities and Memory, Oxford, accessed 31 March 2025, https://citiesandmemory.com/heritage/.
[18] Ibid.
[19] Pınar Yelmi, Hüseyin Kuşcu, and Asım Evren Yantaç, "Towards a Sustainable Crowdsourced Sound Heritage Archive by Public Participation: The Soundsslike Project" (9th Nordic Conference on Human-Computer Interaction, Gothenburg, Sweden, Association for Computing Machinery, New York, 2016); Yelmi, "Protecting Contemporary Cultural Soundscapes as Intangible Cultural Heritage: Sounds of Istanbul."
[20] "The Soundscape of Istanbul Collection - About Collection," Koç University, accessed 31 March 2025, https://librarydigitalcollections.ku.edu.tr/en/collection/soundscape-of-istanbul/.
[21] Ibid.; Yelmi, Kuşcu, and Yantaç, "Short Towards a Sustainable Crowdsourced Sound Heritage Archive by Public Participation: The Soundsslike Project."
[22] Yelmi, Kuşcu, and Yantaç, "Short Towards a Sustainable Crowdsourced Sound Heritage Archive by Public Participation: The Soundsslike Project."
[23] Rhiannon Graybill, "'Hear and Give Ear!': The Soundscape of Jeremiah," *Journal for the Study of the Old Testament* 40, no. 4 (2016): 471, https://doi.org/10.1177/0309089216628414; Schafer, *The Soundscape: our sonic environment and the tuning of the world*, 8.
[24] Schafer, *The Soundscape: our sonic environment and the tuning of the world*, 8.
[25] Ibid.

reconstructions of clothing, three types of sources should be used for historical reconstructions, namely textual, visual and material sources.[26]

Looking at some pioneering works that took up the challenge of reconstructing the sounds of the past, a combination of various types of sources and methodologies being used can be seen. In the fields of acoustical heritage and archaeological acoustics, several types of sources are often brought together in digital reconstructions of specific buildings. For example, Marina Sender et al. based their reconstruction of the Jeromite monastery of Santa Maria de la Murta in Spain on architectural remains, archival materials, and auralisation techniques and recreated the acoustic experience of liturgical events in this church.[27] Similarly, in their study of the Odeon of Pompei, Berardi et al. combined archaeological data, in-situ acoustic measurements and 3D modelling to make a simulation of the original acoustics of the theatre and analysing its changes throughout history.[28]

As these projects show, acoustic research has long concentrated on indoor environments, often using static acoustic simulations based on 3D models of architectural spaces. However, with recent advances in computational power and the development of novel simulation techniques, several innovative studies have emerged extending this type of research to the outdoor environment. This shift has also opened up new avenues for urban soundscape reconstructions. Samuele Briatore, for example, tried to recreate the soundscape of a Baroque festival in Rome using textual sources, such as ceremony records and travel diaries, and visual sources, such as prints.[29] The viewer travels a set route on the print, with different sounds sounding louder or quieter depending on the position of the viewer on the print.[30]

Influential in this regard is also the “Bretez” project, led by Mylène Pardoen.[31] In this project, teams from engineering sciences and humanities combined forces to recreate the soundscape of an area in eighteenth-century Paris. In diverse archival material, such as diaries, police reports, historical

---

[26] Jane Malcolm-Davies, "Structuring Reconstructions: Recognising the Advantages of Interdisciplinary Data in Methodical Research," *Heritage Science* 11, no. 1 (2023): 6, https://doi.org/10.1186/s40494-023-00982-9.
[27] Marina Sender Contell et al., "Virtual acoustic reconstruction of a lost church: application to an Order of Saint Jerome monastery in Alzira, Spain," *Journal of Building Performance Simulation* 11, no. 3 (2018), https://doi.org/10.1080/19401493.2017.1340975.
[28] Umberto Berardi, Gino Iannace, and Luigi Maffei, "Virtual reconstruction of the historical acoustics of the Odeon of Pompeii," *Journal of Cultural Heritage* 19 (2016), https://doi.org/10.1016/j.culher.2015.12.004.
[29] Samuele Briatore, "Immaginare i suoni. Ricostruzione del paesaggio sonoro della festa barocca," *Arti dello Spettacolo / Performing Arts* 3, no. 3 (2017).
[30] Ibid., 91-92, 98-99.
[31] "Bretez," accessed 10 April 2025, https://sites.google.com/site/louisbretez/accueil; Mylène Pardoen, "Projet Bretez: une pincée de son dans l’Histoire.," *Digital Studies/Le champ numérique* 9, no. 1 (2019), https://doi.org/10.16995/dscn.350; "Soundscape Archaeology: a visit to Paris in the mid 18th/early 19th century," Fondation Napoléon, 2020, accessed 10 April 2025, https://www.napoleon.org/en/history-of-the-two-empires/videos/soundscape-archaeology-a-visit-to-paris-in-the-mid-18th-early-19th-century/; "Restoring the Historical Sound of Paris," Hypotheses, 2016, accessed 10 April 2025, https://sms.hypotheses.org/8560.

cartography and paintings, they searched for clues about how the area could have sounded in that period. The different sounds were then recorded and added to a 3D model of the selected area, for which they made use of a game engine called UDK. The result was an interactive environment, in which users could stroll around in the Parisian streets and encounter a dynamic soundscape that reflects the historical auditory environment of the place in the eighteenth century.[32]

A similar study was conducted by Firat et al.[33] The authors used the video game engine Unreal Engine to create a virtual reconstruction of a marketplace in eighteenth-century Naples, in which the user could virtually walk around the square and could hear the dynamic soundscape of the marketplace constructed with various self-made recordings. Firat et al. based their reconstruction on both textual and visual sources. In this research, they also touched upon the problem of the challenging and time-consuming nature of the process of collecting sources for such reconstructions. To address that problem, a keyword gazetteer was used to recognise specific entities in texts, a common method within natural language processing.[34]

Very recently, the University of Antwerp launched a project called *Hearing the Past: Reconstructing the Aural Heritage of Antwerp in the 19th Century*.[35] The project wants to reconstruct the sonic history of Antwerp's inner districts during the nineteenth century and enhance the public engagement with the multisensory history of the city.[36] With the project, attention will be drawn to the importance of sounds as vital components of urban environments and seeing them as elements that merit academic interest and conservation efforts.[37] For the reconstructions, an innovative methodological approach will be used, with artificial intelligence-based natural language processing methods to mine historical texts for sonic data.[38]

As the interest in soundscapes grows, innovative ways of data collection emerge, like the previous two projects have shown. While both of these projects have focused on combining machine learning with textual sources, it might be interesting to look if something similar is possible for visual sources. As Yongho Kim rightly states, images possess significant value as artefacts, since they contain information that cannot

[32] Pardoen, "Restoring the Historical Sound of Paris."; Pardoen, "Projet Bretez: une pincée de son dans l'Histoire.."
[33] Hasan Baran Firat, Luigi Maffei, and Massimiliano Masullo, "Digital Humanities in the Historical Soundscape Research: Sounds of 18th Century Naples" (The Acoustics of Ancient Theatres, Verona, Italy, 2022).
[34] Ibid.
[35] "Hearing the Past: Reconstructing the Aural Heritage of Antwerp in the 19th Century," University of Antwerp, accessed 10 April 2025, https://www.uantwerpen.be/en/projects/hearing-the-past/about/.
[36] Ibid.
[37] Ibid.
[38] Ibid.

be adequately conveyed through textual description alone.[39] Due to several factors such as restricted infrastructure and technical feasibility, up until now the use of visual sources in historical soundscape research was largely limited to manual analysis.

In this context, object recognition, a branch within computer vision in which objects in images or videos are automatically recognised and classified, can be a useful tool for analysing large data sets. With the development of deep learning, and more specifically the emergence of convolutional neural networks (CNNs), the field of computer vision is changing at a rapid pace, with developments of more precise and resilient object detection models as a consequence.[40] Thanks to a technique called transfer learning, in which existing object detection models that are trained on large datasets for a specific task get reused and trained with a smaller dataset for a different task, the use of object detection is already widespread in many sectors, from detecting anomalies in medical images to the detection of objects in self-driving cars.

In contrast, the use of object detection in heritage remains relatively limited. So far, researchers in the fields of conservation, archaeology and art sciences have been among the first to explore the potential of object detection. For example, Ergün Hatir et al. used this technique to develop a model that automatically recognises stone damage, such as cracks and plant growth, in Hittite open-air sanctuaries.[41] Fabrice Monna et al. deployed deep learning on satellite images to recognise traditional houses on the Indonesian island of Sumba. In their study, the authors tried different object detection models and evaluated various techniques for data augmentation.[42] Another example is the work by David Kadish et al., who used object detection to recognise objects in paintings and prints. They fine-tuned Faster R-CNN based on the existing COCO (Microsoft Common Objects in Context) dataset that was modified using AdaIn style transfer.[43]

---

[39] Yongho Kim, Chanjong Im, and Thomas Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," *ACM Journal on Computing and Cultural Heritage* (2024): 1, https://doi.org/10.1145/3699963.
[40] Mupparaju Sohan, Thotakura SaiRam, and Ch. Venkata RamiReddy, "A Review on YOLOv8 and Its Advancements," in *Data Intelligence and Cognitive Informatics. Proceedings of ICDICI 2023.*, ed. I. Jeena Jacob, Selwyn Piramuthu, and Przemyslaw Falkowski-Gilski, Algorithms for Intelligent Systems (Singapore: Springer Singapore, 2024), 1-2.
[41] Ergün Hatir et al., "The deep learning method applied to the detection and mapping of stone deterioration in open-air sanctuaries of the Hittite period in Anatolia," *Journal of Cultural Heritage* 51 (2021), https://doi.org/10.1016/j.culher.2021.07.004.
[42] Fabrice Monna et al., "Deep learning to detect built cultural heritage from satellite imagery. - Spatial distribution and size of vernacular houses in Sumba, Indonesia," ibid.52, https://doi.org/10.1016/j.culher.2021.10.004.
[43] David Kadish, Sebastian Risi, and Anders Sundess Løvlie, "Improving Object Detection in Art Images Using Only Style Transfer," *arXiv (Cornell University)* (2021), https://doi.org/10.48550/arxiv.2102.06529.

In 2022, Odeuropa launched the *Odeuropa Challenge on Olfactory Object Recognition*.[44] This challenge aimed to apply object detection on historical artworks to identify smell-related objects, such as flowers, food, and smoke. Participants were provided with a dataset of 2647 artworks annotated with 20120 bounding boxes and used these to train different state-of-the-art object detection models.[45] The key difficulties that occurred during this project were artistic style variation within the dataset, the detection of small and occluded objects and historically shifting object appearances.[46] With this initiative, Odeuropa wanted to draw attention to the possibilities of interdisciplinary research that uses computational methods.[47]

Very recently, in October 2024, Yongho Kim et al. published innovative research on object detection within digital humanities.[48] Kim et al. wanted to train an object detection model on nineteenth-century children's and youth literature but encountered the barrier of limited annotated data, an issue that is quite common for researcher in this field. To address this problem, the authors proposed an interesting method. Many of the current popular object detection models are trained on the pre-annotated COCO dataset. To leverage this existing data, the authors used CycleGAN to transform the style of these images to the style of illustrations from children's literature, thereby reducing the domain gap and reducing the need for manual annotation. Moreover, they used Pseudo-Labels, a semi-supervised learning technique, which automatically generates annotations based on a small training set. The authors concluded that the combination of these techniques greatly improved the performance of the object detection model.[49]

Research combining machine learning and historical photographs is still scarce. Melvin Wevers applied transfer learning to Places-365, a scene detection model trained on contemporary data, in order to use it on the *De Boer* Collection, a historical press photo collection.[50] In a recent master's thesis, Anil Poudel tested and compared different deep learning networks for facial recognition on historical photographs.[51] But the specific combination of object detection and historical photographs, seems to be non-existent. Given the growing interest in urban soundscapes and the rapidly evolving field of machine learning, I want to provide an initial impetus to fill this gap, namely the lack of knowledge about the performance of object detection on historical photographs. With this pilot study, I want to explore how existing object detection

---

[44] Mathias Zinnen et al., "ODOR: The ICPR2022 ODeuropa Challenge on Olfactory Object Recognition" (26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022).
[45] Ibid.
[46] Ibid.
[47] Ibid.
[48] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling."
[49] Ibid.
[50] Melvin Wevers, "Scene Detection in De Boer Historical Photo Collection," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, ed. Ana Paula Rocha, Luc Steels, and Jaap Van den Herik (Cham: Springer, 2021).
[51] Anil Poudel, "Face recognition on historical photographs" (Uppsala University, 2021).

models perform on historical photographs from an urban context, how transfer learning can be applied to develop a model that can be used within soundscape research and how style transfer with the help of CycleGAN can be used as a tool to alleviate manual annotation work.

This study will seek to answer the central research question: “How can machine learning be used to detect sound sources in historical photographs?” This research question will break down into several sub-questions, which will be the focus of the different chapters. Chapter two outlines the methodology, including a detailed description of the model selection, the data sources, the annotation process, and the training setup. Chapter three will address the first sub-question: “How do existing object detection models, such as YOLOv8, Faster R-CNN and RetinaNet, perform ‘out of the box’ on historical and modern photographs?” In this chapter these three models will be benchmarked on both historical and modern photographs, without training them on this new data. In chapter four, the following sub-question will be central: “How does training an existing class in the model, such as trains/trams, using historical images affect the accuracy of the model?” Here, the best performing model will be fine-tuned by training it on a dataset of historical photographs containing trams. Fine-tuning is a concept that falls under the umbrella of transfer learning in which a pretrained model is further trained on a new dataset, helping it overcome the performance gap caused by visual differences between modern and historical objects.[52] Since the classes in the existing object detection models are limited, they do not always align with the research interests.[53] That is why in chapter five, the following research question will be investigated: “How can an object detection model be extended to include a new class, such as carriages, and how well does the model perform on this new class?” A new class, carriages, will be added here by making use of transfer learning. Chapter six addresses the final sub-question: “Can CycleGAN be used to alleviate annotation work for historical images by transforming existing modern annotated images to a historical style?” Here, CycleGAN will be explored as a method to generate synthetic historical-style images from modern ones, aiming to augment the training data and thereby reducing manual annotation efforts. The penultimate chapter will be used to discuss the results and the potential of this methodology for soundscape and other research. Finally, the conclusion summarises the findings and suggests directions for future research.

[52] Wevers, "Scene Detection in De Boer Historical Photo Collection," 601; "Fine-tuning," Ultralytics, accessed 10 April 2025, https://www.ultralytics.com/glossary/fine-tuning.
[53] Wevers, "Scene Detection in De Boer Historical Photo Collection," 601.

2. Methodology

In this chapter, the methodological approach used to investigate how object detection performs on historical photographs will be discussed. The aim of this research is to propose a workflow for automatically detecting sound sources in historical photographs for sound related research. Therefore, explaining the methodology in as much detail as possible is key to ensure that the workflow can easily be reproduced in future research.

To establish a baseline for determining which object detection model is best suited to train on historical data, a benchmarking study was first conducted to review three popular object detection models, in particular YOLOv8, Faster R-CNN, and RetinaNet, using two datasets: one containing historical photographs in an urban context and a comparable dataset with modern photographs. This allowed for an initial evaluation of how these base models perform on historical images and ensured that the best performing model could be chosen to further refine.

Since the datasets for training were quite small, it was chosen to make use of a common technique called transfer learning. Transfer learning is the general term used to signify the process of using a model pretrained on a large dataset as a starting point for further training with a smaller dataset.[54] With this, the model can transfer knowledge from a previous task, in this case object detection on the COCO dataset, to a different but related task, here the detection of objects on historical photographs. A major advantage of transfer learning is that it avoids training a model from scratch, reducing both training time as well as the amount of data needed for training.[55] One common method of transfer learning is fine-tuning. For fine-tuning, a pretrained model that was trained on a broad dataset is picked, since these models have learnt to recognise general features from their initial training data.[56] During the process of fine-tuning, the weights of the model are adjusted based on the new dataset that it gets trained on. The initial layers of the network, which are responsible for learning general features, are often kept “frozen”, so that their weights are not updated, and the later, more task-specific layers are retrained. However, this freezing is not mandatory.[57]

In this study, the best performing model was first fine-tuned on a “known” class, “train”, meaning that this class was also present in the COCO dataset on which the model was originally trained. Since the model

[54] Uday Kulkarni et al., "Classification of Cultural Heritage Sites Using Transfer Learning" (IEEE Fifth International Conference on Multimedia Big Data (BigMM), Signapore, 2019); Poudel, "Face recognition on historical photographs," 17-18.
[55] Wevers, "Scene Detection in De Boer Historical Photo Collection," 604; "Fine-tuning."
[56] "Fine-tuning."
[57] Ibid.

had been pretrained on this COCO dataset, this step examined how well the model could be fine-tuned on a more specialised dataset of historical photographs containing trams in an urban context. The rationale behind this choice was that, due to the potential similarities between modern and historical representations, the model might require less data to achieve acceptable performance when adapting to the historical domain. To further test the adaptability of the model, a new class, “carriages”, was introduced, also by making use of fine-tuning. This step served as a first exploration of how well custom classes can be integrated into the model, which will be necessary for future soundscape-related research. The training was conducted using both a base dataset and an augmented one, to see how much difference the augmentation made. Finally, to address the challenge of limited historical training data and the time-consuming process of annotating the training data, CycleGAN was put to the test to generate “fake” historical images from modern photographs and vice versa.

The following sections will go into deeper detail on the different object detection models, on the used sources, the annotation process, the augmentation and preprocessing steps, the final datasets, the training details and finally the evaluation metrics used throughout this research.

2.1. Object detection models and CycleGAN

In the field of machine learning, object detection has seen rapid advancements the last few years thanks to the continuous development of more efficient and accurate algorithms.[58] Different models with different architectures and purposes are being released regularly. Three of the most popular object detection models, YOLOv8, Faster R-CNN and RetinaNet, were chosen for this study. These three models were picked because of their popularity, meaning that these models are well documented, open source and used throughout different sectors. Moreover, object detection models are divided into two categories, namely single-stage and two-stage detectors. Two-stage detectors, like Faster R-CNN, involve separate steps for region proposal and object classification. This is in contrast to single-stage detectors, like YOLOv8 and RetinaNet, which perform both tasks simultaneously, which means that it detects objects in an image through a single forward pass of the neural network.[59] Both of these have their own advantages, with single-stage detectors typically being faster, which is useful for real-time applications and two-stage detectors generally being more accurate.[60] Because of their different characteristics, it was made sure that both types of models were included in this research. In the following parts, these three

[58] Momina Liaqat Ali and Zhou Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," *Computers* 13, 12 (2024): 1, https://doi.org/10.3390/computers13120336.
[59] Ibid., 4.
[60] Bsher Karbouj, Garabet A. Topalian-Rivas, and Jörg Krüger, "Comparative Performance Evaluation of One-Stage and Two-Stage Object Detectors for Screw Head Detection and Classification in Disassembly Processes," *Procedia CIRP* 122 (2024): 528, https://doi.org/10.1016/j.procir.2024.01.077.

models will be discussed. The intention is not to describe the complex architectures of all these models, but rather to provide basic information about the models and why these were chosen.

2.1.1. YOLOv8

Among the most popular object detection models available today is YOLO, short for "You Only Look Once". YOLO is actually a family of computer vision models, with different versions released over time. The original YOLO model was developed in 2015 and presented in 2016 by Joseph Redmon et al.[61] They revolutionised the field of object detection by introducing a model that used a grid-based approach to predict bounding boxes and class probabilities at the same time.[62] This meant that YOLO was able to accomplish detecting objects in images by passing the image only one time through the network, hence why the name "You Only Look Once".[63] YOLO quickly gained popularity, due to its efficiency which made it highly suitable for real-time applications. Since the original model, YOLO has had various evolutions, each focusing on improving the balance between speed and accuracy.[64] The latest version, YOLOv12, was released on February 18th, 2025. The models after YOLOv3 are developed by other authors than Redmon et al. and they each have varying goals based on the preference of the authors.[65]

For this thesis, YOLOv8 was chosen due to its compatibility with Roboflow, the platform used for annotating images, and because of its extensive documentation and prebuilt Google Colab notebooks.[66] YOLOv8 is developed by Ultralytics, a team that also made YOLOv5. It was released on January 10th, 2023.[67] The framework can be used to perform different computer vision tasks, like detection, segmentation, classification, and pose estimation. Each of these tasks has pretrained models which can be used to work upon. Different scaled versions of the model are provided by Ultralytics, ranging from YOLOv8n (nano), made for very fast inference with lower computational demands, to YOLOv8x (extra large), which offers higher accuracy at the cost of increased computational demands. In this thesis, YOLOv8l was selected due to the promised strong performance while maintaining a reasonable inference time.

---

[61] Joseph Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," *arXiv (Cornell University)* (2015), https://doi.org/10.48550/arXiv.1506.02640.
[62] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," 7-8.
[63] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction* 5, no. 4 (2023): 1685, https://doi.org/10.3390/make5040083.
[64] Ibid., 1708.
[65] "What is YOLO? The Ultimate Guide [2025]," Roboflow, 2025, accessed 2 April 2025, https://blog.roboflow.com/guide-to-yolo-models/#yolov12.
[66] Ultralytics YOLOv8.
[67] Sohan, SaiRam, and RamiReddy, "A Review on YOLOv8 and Its Advancements," 4.

2.1.2. RetinaNet

The second model used for benchmarking is RetinaNet, an object detection model introduced in 2017 in a paper by Tsung-Yi Lin et al.[68] Just like YOLO, RetinaNet is also a single-stage detector. It was developed with the aim to overcome the usual reduced accuracy in single-stage detectors and thus creating a balance between speed and accuracy.[69] This was done by incorporating a Feature Pyramid Network (FPN) for multi-scale feature extraction, meaning that it enhances the ability to detect objects at different scales, and by employing a Focal Loss function, which tackles the class imbalance problem during training by modifying the standard cross-entropy loss function to down-weight the loss assigned to well-classified examples.[70] This means that it gives priority to harder-to-detect objects during training, which helps overcoming the imbalance between foreground and background classes.[71] For this research, where historical photographs are often of worse quality containing noise and smaller or occluded objects, RetinaNet was chosen precisely because of its characteristics to perform good on harder-to-detect objects. In this case, an improved RetinaNet model with a ResNet-50-FPN backbone was used.

2.1.3. Faster R-CNN

The last model selected for this research is Faster R-CNN, developed by Shaoqin Ren et al. in 2015, as a successor to R-CNN (Region-based Convolutional Neural Network) and Fast-RCNN.[72] Contrary to the previously discussed models, Faster R-CNN is a two-stage detector, meaning that it first generates region proposals and then classifies and refines these proposals. This additional step allows for more precise detection, but this is at the expense of computational speed.[73] The key innovation of Faster R-CNN is the Region Proposal Network (RPN), which generates proposals where an object may be located in a single forward pass, leading to a significant reduction in computation time and an improvement in accuracy compared to R-CNN and Fast-RCNN.[74] As these two-stage detectors are often seen as more accurate than their single-stage counterparts, it was logical to include one in this thesis and measure how well it

---

[68] Tsung-Yi Lin et al., "Focal Loss for Dense Object Detection," *arXiv (Cornell University)* (2017), https://doi.org/10.48550/arXiv.1708.02002.
[69] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," 6.
[70] Ibid.; "RetinaNet: Single-Stage Object Detector with Accuracy Focus," viso.ai, 2024, accessed 3 April 2025, https://viso.ai/deep-learning/retinanet/.
[71] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," 6.
[72] Shaoqing Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *arXiv (Cornell University)* (2015), https://doi.org/10.48550/arXiv.1506.01497.
[73] Monna et al., "Deep learning to detect built cultural heritage from satellite imagery. - Spatial distribution and size of vernacular houses in Sumba, Indonesia," 174.
[74] Sohan, SaiRam, and RamiReddy, "A Review on YOLOv8 and Its Advancements," 3.

performs on historical photographs. In this research, a pretrained and improved Faster R-CNN with a ResNet-50-FPN backbone was used.

2.1.4. CycleGAN

In addition to the object detection models, CycleGAN was used in this research to address the challenge of data scarcity. As its name suggests, the basis of CycleGAN are GANs or Generative Adversarial Networks. GANs are a class of machine learning frameworks introduced in 2014 in a paper by Ian Goodfellow et al.[75] These frameworks are designed to generate new data based on a training dataset. The output data can take on many forms, such as images, videos or texts.[76] At the core of GANs are two neural networks, namely the Generator and the Discriminator, which are trained simultaneously through adversarial training.[77] The goal of the Generator is to produce new data that resembles the input data as close as possible. The Discriminator on the other hand takes the input data and the new data produced by the Generator and attempts to distinguish between the two and outputs the probability that the data is real.[78] By training the GAN, a sort of collaboration is formed. The Generator wants to produce data that the Discriminator cannot distinguish from real data while the Discriminator tries to get better at discerning between generated data from authentic data. The process thus makes it more challenging for the Discriminator at distinguishing while at the same time encouraging the Generator to produce higher quality data.

CycleGAN takes this a few steps further. This can be illustrated based on the core of this thesis, namely historical and modern pictures. The dataset is split into two domains, domain A containing modern, contemporary, photographs and domain B existing of historical photographs. Then there are two Generators and two Discriminators. On the one hand there is Generator B, further called GB. This Generator takes a real image from domain A (a modern one) and transforms it into a fake image in domain B (a historical image). The second Generator, Generator A (further GA), does the inverse of this process. It takes a real historical image and translates it into a fake modern image. On the other hand, there are the Discriminators. Discriminator B (further DB) is in charge of deciding whether an image from domain B, so a real historical image or a modern turned historical image, is real or fake. Discriminator A (further

---

[75] Ian Goodfellow et al., "Generative Adversarial Networks," *arXiv (Cornell University)* (2014), https://doi.org/10.48550/arxiv.1406.2661.
[76] "Guide to Generative Adversarial Networks (GANs) in 2025," viso.ai, updated 1 October 2024, 2024, accessed 24 December 2024, https://viso.ai/deep-learning/generative-adversarial-networks-gan/.
[77] "Generative Adversarial Networks," Medium, updated 30 October 2023, 2023, accessed 24 December 2024, https://medium.com/@marcodelpra/generative-adversarial-networks-dba10e1b4424.
[78] Ibid.

DA), does the opposite, it discerns between real or fake modern images.[79] As illustrated in Figure 1, each Generator gets used a second time, which is the crux of CycleGAN. This process is called cycle consistency, which means that an image that belongs to domain A is transformed into a fake image of domain B by GB and then it is transformed back into a reconstructed version of the original, in domain A, by GA. The resulting reconstructed image should resemble its original counterpart as close as possible.

This leads to another important part of the architecture of CycleGAN, namely the three loss functions that measure how well the model works. First, there is the adversarial loss function, which is fairly straightforward. This function measures how well the Discriminators can tell fake images from real ones, and how successful the Generators are in fooling the Discriminators and thus in producing high quality fake images. The second loss function, called cycle consistency loss, is, as previously stated, the main part of CycleGAN. This loss function ensures that an image looks as close as possible to the original image after passing a whole cycle, so after it has been translated from one domain to another and then back to the original. It encourages the model to alter only the style of the image while preserving the content of it. Finally, there is the identity loss function, which is used to make sure that the Generator does not change images that it should not change. If GB for example gets presented an historical image, it should just return the same historical image. The identity loss thus ensures that there is a close alignment between input and output when images from the target domain were fed into the generator, which prevents transformations that would alter the overall colour tone of the photographs.[80]

[79] Djarot Hindarto, "Revolution in Image Data Collection: CycleGAN as a Dataset Generator," *Sinkron : Jurnal Dan Penelitian Teknik Informatika* 8, no. 1 (2024): 449, https://doi.org/10.33395/sinkron.v9i1.13211.
[80] Jun-Yan Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *arXiv (Cornell University)* (2017): 8, https://doi.org/10.48550/arxiv.1703.10593.

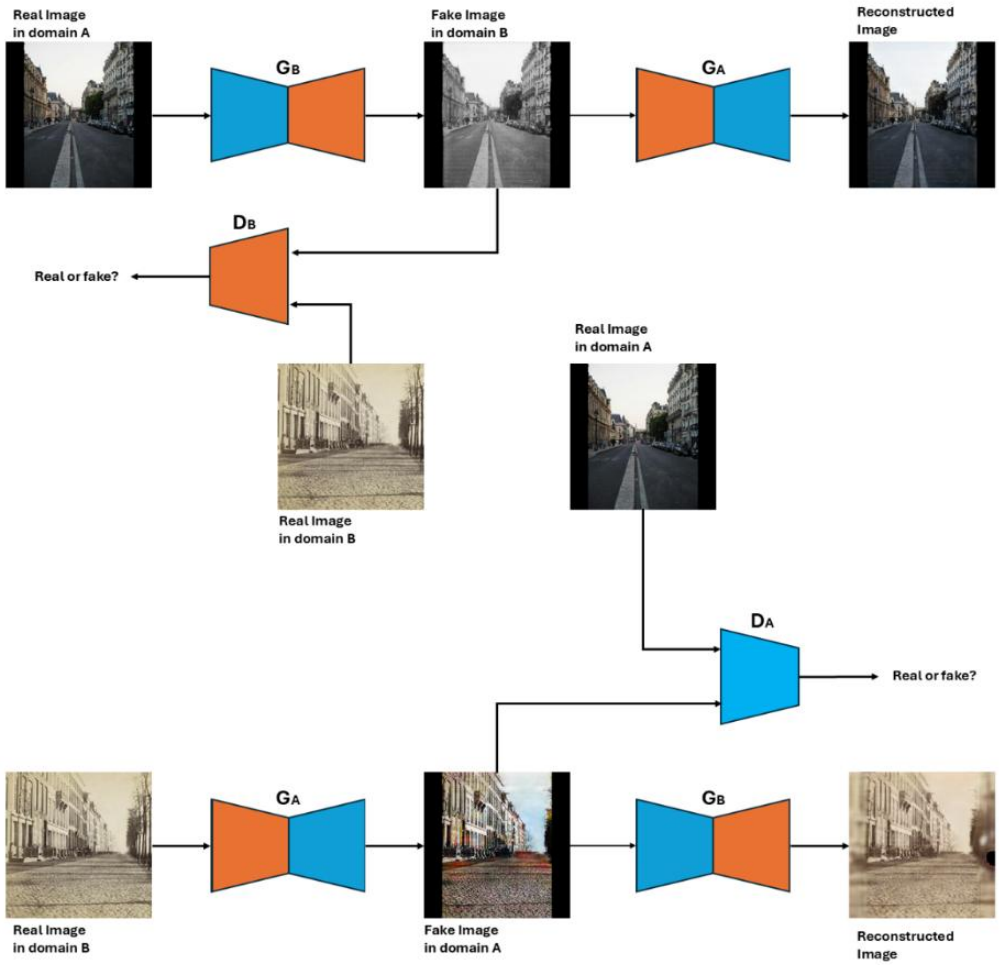


Figure 1. CycleGAN illustrated. Figure made by author, inspired by figure 3 in Djarot Hindarto, "Revolution in Image Data Collection: CycleGAN as a Dataset Generator," Sinkron : Jurnal Dan Penelitian Teknik Informatika 8, no. 1 (2024): 449, https://doi.org /10.33395/sinkron.v9i1.13211. (Sources of the images: Real Image in domain A: Peeters, Florian. A city street with cars parked on both sides*. Photograph. Unsplash. June 19, 2023. https://unsplash.com/photos/a-city-street-with-cars-parked-on-both-sides-qBWhIm1JPB4. Real Image in domain B:* Boompjes Rotterdam*. Photograph. Europeana (Rijksmuseum). 1860-1880. https://www.europeana.eu/nl/item/90402/RP_F_F12132.)*

2.2. Sources

A popular saying in the field of machine learning is “garbage in, garbage out”, meaning that a machine learning model is only as good as the data that it learns from. The first step in the creation of a good dataset for object detection, therefore, is finding suitable pictures from which the model can learn. A dataset must be diverse so the model can adapt to different situations and it also has to be representative of the tasks the model has to perform. Given that this research focuses on developing an object detection model for detecting sound sources in an urban context, the dataset was carefully curated to be representative of urban environments, with images depicting cities, busy streets, marketplaces, …

An important consideration in this study is the legal framework surrounding data collection. The legal landscape for data collection in machine learning is evolving at a rapid pace. In a European context, the

Directive 2019/790/EU on copyright in the Digital Single Market (DSM Directive) and the AI Act are the most important frameworks. The DSM introduced exceptions for the so called “data mining” on lawfully accessible content unless the rightsholders explicitly have opted out.[81] The AI Act has further imposed guidelines on AI development, with an emphasis on transparency and ethical AI training practices.[82] However, scientific research is exempted from this AI act. Nonetheless, for this research only photographs in the public domain or licensed under terms that allow unrestricted use, such as Creative Commons CC0, were used.

2.2.1. Historical photographs

The primary source for historical photographs was Europeana, an online cultural heritage platform launched in 2008 as part of the European Commission’s i2010 strategy.[83] It is operated by the Europeana Foundation and provides access to over 60 million digital objects from over 2000 different cultural institutions, making it the largest digital cultural heritage aggregator of Europe. Europeana emerged out of Europe’s policy to preserve and showcase Europe’s cultural diversity and foster an identity.[84] Europeana makes it easy to browse through very diverse sources of photographs from across Europe, thus greatly broadening the scope of data collection. In addition to Europeana, several other archives were consulted to increase the variety in the dataset. These were the FelixArchief, the website ErfgoedBrugge, the Erfgoedbank Brussel and the online archive of the Library of Congress.[85]

Following search terms were used to find images: Amsterdam, Antwerp, Berlin, Birmingham, Bruges, Brussels, carriage, Den Haag, Ghent, Groningen, koets, Kortrijk, Liverpool, London, market, Mechelen,

[81] "Artificial intelligence and copyright: use of generative AI tools to develop new content," European Commission, 2024, accessed 4 April 2025, https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_en; "AI ‘opt-outs’: should cultural heritage institutions (dis)allow the mining of cultural heritage data?," Europeana Pro, 2024, accessed 4 April 2025, https://pro.europeana.eu/post/ai-opt-outs-should-cultural-heritage-institutions-dis-allow-the-mining-of-cultural-heritage-data.
[82] "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," ed. European Union and European Council (2024). http://data.europa.eu/eli/reg/2024/1689/oj.
[83] "Europeana," Europeana, accessed 10 October 2024, https://www.europeana.eu/nl.
[84] Carlotta Capurro and Gertjan Plets, "Europeana, EDM, and the Europeanisation of Cultural Heritage Institutions," *Digital Culture & Society* 6, no. 2 (2020): 165-66, https://doi.org/10.14361/dcs-2020-0209.
[85] "FelixArchief," Stad Antwerpen, accessed 10 October 2024, https://felixarchief.antwerpen.be/; "Erfgoed Brugge," Erfgoed Brugge, accessed 10 October 2024, https://erfgoedbrugge.be/; "Erfgoedbank Brussel," Erfgoedbank Brussel, accessed 10 October 2024, https://erfgoedbankbrussel.be/; "Library of Congress," Library of Congress, accessed 10 October 2024, https://www.loc.gov/.

Munich, Oostende, Ostend, Paris, River, Rome, Rotterdam, Scheldt, Sint-Niklaas, tram, Utrecht, Venice, Vienna.

2.2.2. Modern photographs

Next to the historical dataset, a dataset consisting of modern photographs was compiled, using photographs from Unsplash.[86] This website is owned by Getty Images and provides a wide array of free to use photographs. The pictures were collected by using the same search terms as for the historical photographs, and it was made sure that the collected photographs were a good modern counterpart of the historical ones. This ensured that the benchmarking would be fair.

2.3. Annotation

The annotation process for this research was done using Roboflow, an online platform that facilitates different stages of developing a model, from dataset management, annotation and preprocessing to the deployment of models for computer vision tasks.[87] Roboflow has different pricing plans, including a free tier with limited features and paid plans that provide access to more options like enhanced dataset augmentation. They also have a free plan for academics, where they get access to a few more resources for data augmentation and training. This academic plan was used for this research.

The annotation process consisted of drawing bounding boxes around the objects or classes that needed to be detected. For the benchmarking test, these classes were “person”, “train”, “boat” and “horse”, four classes that are part of the 80 COCO classes, and therefore already known to the model. These classes were chosen because they could be linked to sources that produce sound and because they prominently feature in many historical photographs. The class “train” was then used to further refine the model and a custom class, “carriage”, was added. This meant that in total, five different sorts of objects needed to be annotated on the photographs. The bounding boxes needed to be drawn with attention, since the quality of this would determine the performance of the model. Thus, effort was made to ensure that the boxes were as tight as possible around the objects and that similar objects were annotated uniformly across all images. After the annotation, the different datasets were inspected one final time to make sure that all objects were annotated and the bounding boxes were drawn correctly.

To accelerate the annotation process, Roboflow offers an interesting feature called “Auto Label”. With this tool, the user can choose one of Robfoflow’s pretrained object detection models or upload an own

[86] "Unsplash," Unsplash, accessed 10 October 2024, https://unsplash.com/.
[87] "Roboflow," Roboflow, accessed 14 November 2024, https://roboflow.com/.

model to automatically create bounding boxes and label data, reducing the manual annotation workload. Roboflow claims that the tool can reduce labelling time by more than 50%.[88] For the annotation of the "train" and "carriage" classes, this tool was tested by using the "Grounding DINO" model and providing it with the class names, a description of the classes and a confidence threshold of 25%. While this approach provided a useful starting point, it was not entirely reliable. As can be seen in Figure 2, a lot of objects were missed, incorrectly classified, or the bounding boxes were bad, which meant that it needed many manual corrections. As a result, most of the annotation process was completed manually to ensure high-quality training and test data. Experimentation with prompt engineering for future research could be interesting to see if Grounding DINO works better with well-crafted prompts.

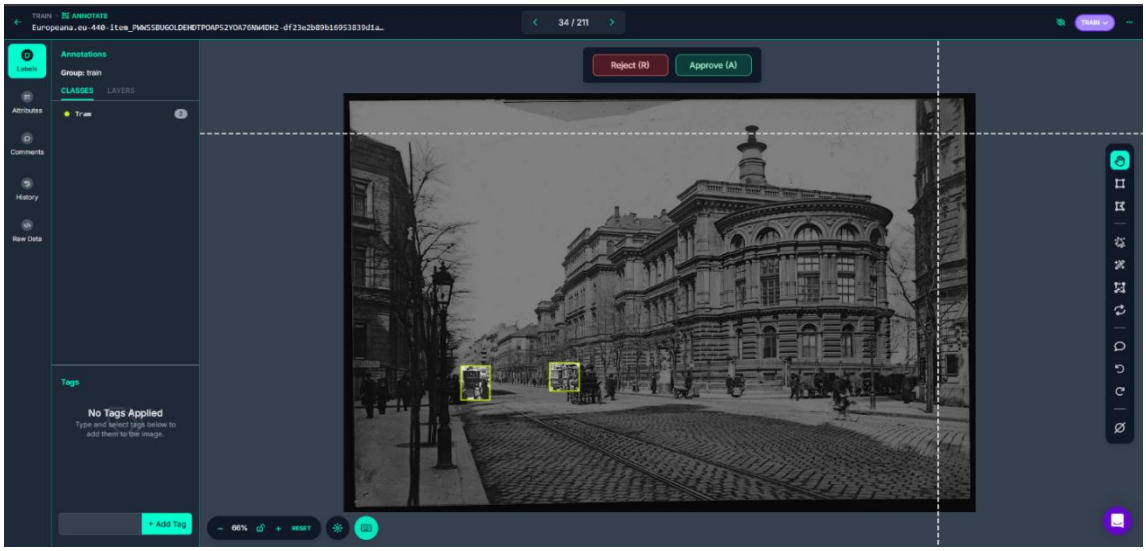


Figure 2. The annotation screen in Roboflow. This picture shows the output of Roboflow's Auto Label. It can be seen that the bounding box on the left is not tight enough. (Source image: Budapest. Klinika Üllöi út (1877). *Photograph. Europeana (Deutsche Fotothek). 1877. https://www.europeana.eu/nl/item/440/item_PWWSSBUGOLDEHDTPOAP52YOA76NW4DH2.)*

2.4. Augmentation

Training a machine learning model requires a large amount of data, which can be challenging to obtain when working independently due to the time-consuming nature of annotation. A simple technique that can help artificially enlarge datasets is called data augmentation.[89] This technique increases the diversity of the training data by applying various transformations to the existing images and helps prevent overfitting. Overfitting occurs when the model adapts too closely to the training data, achieving excellent performance on the training data, but performing significantly worse on unseen data. By applying data augmentation, the number of training images that the model learns from increases, making it harder for

[88] "Automatically Label Image Data," Roboflow, accessed 14 November 2024, https://roboflow.com/auto-label.
[89] Monna et al., "Deep learning to detect built cultural heritage from satellite imagery. - Spatial distribution and size of vernacular houses in Sumba, Indonesia," 175.

the model to overfit to a particular image.[90] At the same time, excessive data augmentation could lead to underfitting. This happens when the model does not perform well on both the training dataset and on unseen data.[91] When there is excessive data augmentation, the model could have difficulties to extract generalisable features.[92]

Roboflow offers the possibility to apply various augmentation techniques, a few of which have been applied in this study. Brightness was adjusted by +/- 15% to simulate a range of lightning conditions, which would make the model more resilient to different conditions. Exposure was modified by +/-10%, which means that the highlights and shadows were affected, assisting the model with being versatile to lighting and camera setting changes.[93] Next, a random Gaussian blur with a radius of 2.5 pixels was added to simulate motion blur or objects not in focus, which are common in historical sources. Saturation was altered by +/- 25%, which means that the vibrancy of the colours in the images was adjusted randomly. Finally, noise was added to 0.1% of the pixels, which helps the model to be more resilient to camera artefacts. These techniques were chosen because these are all factors that show up in real historical photographs. For the values of the augmentations, the original settings of Roboflow were used, which, as can be seen in Figure 3, create only relatively subtle changes.

Figure 3: The different augmentation techniques used in this research. (Source image: De Keyserlei: het Centraal Station, Antwerpen 1910*. Photograph. FelixArchief. 1910.* *https://felixarchief.antwerpen.be/detailpagina?invnr=PB_2005&dtnr=1224_40&dtrecordid=22623&page=1&pageS.)*

[90] Wevers, "Scene Detection in De Boer Historical Photo Collection," 605.
[91] Poudel, "Face recognition on historical photographs," 19-20.
[92] Wevers, "Scene Detection in De Boer Historical Photo Collection," 606.
[93] Brahim Jabir, Noureddine Falir, and Khalid Rahmani, "Accuracy and Efficiency Comparison of Object Detection Open-Source Models," *International Journal of Online and Biomedical Engineering* 15, no. 5 (2021): 173, https://doi.org/10.3991/ijoe.v17i05.21833.

2.5. Preprocessing

After the augmentation, all the annotated images were preprocessed, also by using Roboflow. Preprocessing ensures that the images are converted in a suitable and consistent format for training and testing the different models. The most important step here was the resizing of the images to a uniform resolution. For training of the object detection models, the images were resized to a resolution of 640 x 640 pixels, a common input size for many object detection models. Because not all images were the same size, it was necessary to apply “padding” when resizing. This means that, depending on the original size of the images, a black border got added to the pictures. By doing this, the original aspect ratio was maintained, and the images did not get stretched or cropped. For the training of CycleGAN, the photographs were resized to 256 x 256 pixels due to computational demands.

The images were then split into training, validation and test sets. The default split was 70% training data, 20% validation data, and 10% testing data. However, when data augmentation techniques were used, only the training images were augmented. This caused the relative proportion of each dataset to vary depending on the number of augmented images added to the training set. Finally, the different datasets were exported in the appropriate annotation formats. For the YOLOv8 model, the annotations were converted into the YOLO format and for the Faster R-CNN and RetinaNet models, the COCO format was used. These datasets were then uploaded to Google Drive, to make training with Google Colab go smoother.

2.6. Final datasets

This section provides an overview of the different datasets created and used throughout this research project. First, the two different datasets used for the benchmarking of the three object detection models will be discussed, followed by the different datasets for training the best performing model on the class “train” and the added class “carriage”. Finally, a brief description of the dataset used for CycleGAN will be given.

2.6.1. Benchmarking

For the benchmarking of the different object detection models discussed previously, two separate datasets were curated, one containing historical photographs and one containing modern photographs. Despite the availability of different pre-annotated datasets containing modern images, it was chosen to create a custom dataset to ensure that it reflected the same thematic and contextual characteristics as the historical dataset. In particular, the two datasets mainly consisted of scenes situated in urban environments, thereby providing a consistent basis for comparison between the two datasets and

allowing for a more accurate assessment of model performance across distinct but contextually similar data.

Each dataset contained 300 images and was annotated for the four selected classes. As can be seen in Table 1 and Table 2 below, the datasets were not perfectly balanced in terms of the number of object instances per class. This was caused by the fact that some of the classes, like “person”, just appear more frequently on pictures in this urban context. As Kim et al. state, these “imbalanced datasets” have an impact on the training of a custom model, since strong bias toward the majority class during training could occur, which causes that the model achieves high overall accuracy but fails to predict items in the minority classes.[94] While this issue is less critical in the context of benchmarking, since the models were not retrained on the data, it remains an important consideration. Smaller class sizes are indeed more susceptible to fluctuations in the performance scores.

Modern Dataset		
Class	**Images**	**Instances**
All	300	3169
Person	264	2701
Train	131	148
Boat	63	251
Horse	26	69

Table 1: Modern Dataset used for benchmarking the different object detection models.

Historical Dataset		
Class	**Images**	**Instances**
All	300	3555
Person	270	2896
Train	80	116
Boat	53	227
Horse	137	316

Table 2: Historical Dataset used for benchmarking the different object detection models.

2.6.2. Train

To train the best performing object model on a known class, “train”, corresponding to the COCO class, seven different datasets were created. The first dataset is called D0-Base, which is the base dataset containing a total of 422 pictures of which 295 were used for training. These 295 pictures contained 474 instances, meaning that 474 bounding boxes were drawn around trains/trams. Next, there were five datasets (D1-BRT, D2-EXP, D3-BLR, D4-SAT and D5-NS) all using different augmentation techniques, respectively brightness, exposure, blur, saturation and noise. These five datasets were created to see which of these augmentation techniques got the best results, which is something interesting to know for future research. All the training images were augmented three times. However, as can be seen in Table 3, they do not all contain exactly three times the number of images as the original. This discrepancy is due to the way Roboflow handles augmentation. Rather than duplicating each image exactly three times, Roboflow generates a total number of augmented images that is approximately three times the size of

[94] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 9-10.

the original training set, but the selection process involves random sampling. Dataset D6-Allx3 and dataset D7-Allx7 combined all these augmentation techniques, respectively being augmented three times and seven times. The different sizes of the datasets make it possible to see if increasing the number of training data has an impact on the accuracy of the model.

Name	Augmentation	Total Pictures	Training Pictures	Instances training pictures
D0-Base	No augmentation	422	295	474
D1-BRT	Brightness	1003	876	936
D2-EXP	Exposure	1001	874	935
D3-BLR	Blur	987	860	910
D4-SAT	Saturation	1006	879	940
D5-NS	Noise	1012	885	948
D6-Allx3	All of the above , x3	1012	885	948
D7-Allx7	All of the above, x7	2192	2065	2844

Table 3: The different datasets used to train an object detection model on the class "train".

2.6.3. Carriages

This research also explored training a model on a completely new class, “carriage”. Because the model did not know this class already from its previous training, it was supposed that it needed more training data than the previous “train” class. Two datasets were created. D0-Base was, as the name suggests, the base dataset containing a total of 1159 pictures, of which 821 were used for training the model. These training pictures contained 2768 carriages. A second datasets, D1-Allx7, was created by deploying all the different augmentations discussed in the previous section and augmenting the training pictures seven times. In the free version of Roboflow, the images can be augmented only three times. The version for students and academic research increases this to seven times. When using one of the different paid versions of Roboflow, this number can be even higher.

Name	Augmentation	Total Pictures	Training Pictures	Instances training pictures
D0-Base	No augmentation	1159	821	2768
D1-Allx7	All discussed augmentations, x7	6085	5747	16734

Table 4: The different datasets used to train an object detection model on the class "carriage".

2.6.4. CycleGAN

One of the notable features of CycleGAN is the fact that it does not require paired data to train on. In this context, it means that it can use historical photographs and modern ones with varying scenes and themes. It does not need to have “matching” sets of pictures. What is important however, as stated by Jianwei Bai in his research into CycleGAN and historical Chinese paintings, is the fact that “the diversity of image styles within the dataset is crucial for training the model to adapt to various themes and scenes, which enhances the model’s generalization capabilities, enabling it to excel in the task of style transformation for modern photos.”[95] During data collection, it was made sure that this demand for diversity was fulfilled. Although all photographs came from an urban setting, different kinds of sceneries were used. This means different cities, different scenes within the city and using both close-ups as well as photographs taken from further away. In total, dataset A (modern photographs) consisted of 780 images, dataset B (historical photographs) of 897 images.

2.7. Training

Deep learning projects, such as object recognition models, rely on heavy computation on large datasets, requiring significant computational resources. Graphics processing units (GPUs) are essential for most projects to achieve a reasonable training time, as they are specifically designed to handle the parallel processing demands of deep learning.[96] However, access to powerful GPUs can be expensive, making it not always feasible for individual researchers. Therefore, in this study Google Colab was used.[97] Google Colab is a cloud service developed by Google for disseminating machine learning education and research. It is based on Jupyter Notebooks, which is an open-source and browser-based tool that integrates interpreted languages, libraries and tools for visualisation.[98]

Google Colab is particularly well-suited for machine learning projects like this research, as it provides access to a range of different computational resources like GPUs and TPUs (Tensor Processing Units), without the need for local hardware. One of the key advantages of Google Colab is its integration with Google Drive, which allows users to easily store and access datasets, notebooks, and model outputs.[99] Furthermore, Colab notebooks can be shared with other users, enabling collaborative research and

---

[95] Jianwei Bai, "Ancient Chinese Painting Style Transfer Based on CycleGAN," *Applied and Computational Engineering* 51, no. 1 (2024): 130, https://doi.org/10.54254/2755-2721/51/20241192.
[96] Tiago Carneiro et al., "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications," *IEEE Access* 6 (2018): 61677, https://doi.org/10.1109/ACCESS.2018.2874767.
[97] "Google Colaboratory," Google, accessed 14 November 2024, https://colab.google/#:~:text=Google%20Colaboratory,%2C%20data%20science%2C%20and%20education.
[98] Carneiro et al., "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications," 61678.
[99] Ibid.

reproducibility. While the free version of Colab offers access to a GPU, this access is limited in terms of availability and session duration, which can cause runtime disconnections where the training progress is lost. To ensure stable GPU access during training for this research, an upgrade was made to Google Colab Pro, which provided access to high-performance GPUs and more reliable connections. Google Colab Pro costs €11,19 per month, which gives the user 100 “computational units” to use the GPUs.

2.7.1. Benchmarking

To benchmark the three object detection models, YOLOv8, Faster R-CNN, and RetinaNet, six different Colab notebooks were created, each dedicated to one model and one dataset.[100] Each model was applied in inference mode on the respective datasets, allowing for a direct comparison without additional training. The YOLOv8 model was loaded using the Ultralytics Python package. The pretrained large variant of the model was then selected and executed in the validation mode, a mode used to evaluate a trained model on a validation set to measure its accuracy and generalisation performance.[101] The Faster R-CNN and RetinaNet models were accessed through the torchvision library, using the fasterrcnn_resnet50_fpn_v2 and retinanet_resnet50_fpn_v2 architectures respectively, each initialised with COCO pretrained weights.[102] They were then also set to evaluation mode.

2.7.2. Training the model

As will be discussed in the next chapter in which different object detection models were benchmarked, YOLOv8l performed best on the historical pictures and was further used to train on this historical data. For training two custom models, one for the known class containing images of trains and trams and one for the unknown class containing carriages, a premade YOLOv8 notebook by Roboflow was used.[103] This notebook already contained all the necessary steps to develop a custom model and thus it was only “plug and play” by adding the own data to the script.[104]

---

[100] In the bibliography a Google Colab notebook can be found containing all the used code in this research.
[101] Sohan, SaiRam, and RamiReddy, "A Review on YOLOv8 and Its Advancements," 9.
[102] "fasterrcnn_resnet50_fpn_v2," PyTorch, accessed 10 January 2025, https://pytorch.org/vision/main/models/generated/torchvision.models.detection.fasterrcnn_resnet50_fpn_v2.html#torchvision.models.detection.fasterrcnn_resnet50_fpn_v2; "retinanet_resnet50_fpn_v2," PyTorch, accessed 10 January 2025, https://pytorch.org/vision/0.20/models/generated/torchvision.models.detection.retinanet_resnet50_fpn_v2.html.
[103] "train-yolov8-object-detection-on-custom-dataset.ipynb," Google Colab, accessed 24 January 2025, https://colab.research.google.com/github/roboflow-ai/notebooks/blob/main/notebooks/train-yolov8-object-detection-on-custom-dataset.ipynb.
[104] In the bibliography a Google Colab notebook can be found containing all the used code in this research.

For the training on the known class, the YOLOv8 large model was trained for 100 epochs with a batch size of 16 and an input resolution of 640x640 pixels. Google Colab's best performing GPU, the T4, was used to get the results as quick as possible. For the hyperparameters, namely the batch size, learning rate, weight decay and momentum, the standard settings of the model were used. This meant that the batch size was 16, the initial learning rate was 0.01 with a learning rate final fraction also of 0.01, the momentum was 0.937, and the weight decay was 0.0005.

As stated previously, with fine-tuning the choice can be made to freeze the initial layers of the network which are responsible for learning the more general features.[105] By freezing, the parameters of the layers (weights and biases) do not get updated during training. Freezing layers can be beneficial in different situations, for example when there are limited computational resources, when the new dataset is significantly smaller than the original dataset or when there are similar feature domains between the new dataset and the one the model was trained on.[106] On the other hand, freezing can lead to a slight reduction in the performance of the model, and knowing which layers to freeze is also an experimental process.[107] It was thus chosen to not freeze any layers for this research. Future research could maybe explore layer freezing as an additional variable in optimising object detection on historical images.

In addition to the augmentations applied via Roboflow, YOLOv8's built-in pipeline further transformed the images from epoch to epoch by horizontal flipping, scaling and colour distortions for example. After each epoch, the model was evaluated on the validation set. For the training of the model on a new class, the same configuration and settings were applied. However, the number of training epochs was reduced to 90 due to repeated runtime failures on Colab when attempting to train for 100 epochs. Despite six attempts, the training process consistently disconnected near the final epoch after nearly 8 hours of training which also wasted the bought computational units. As a result, the decision was made to slightly reduce the amount of epochs.

2.7.3. CycleGAN

For training the CycleGAN model, the Google Colab notebook as provided by the original authors of CycleGAN was used, which can be found on GitHub.[108] The L4 GPU was used for training, which is the second-best GPU offered by Google Colab. For the training process, the standard options of the original

---

[105] "Fine-tuning."
[106] "Transfer Learning with Frozen Layers in YOLOv5," Ultralytics, accessed 10 April 2025, https://docs.ultralytics.com/yolov5/tutorials/transfer_learning_with_frozen_layers/.
[107] Ibid.
[108] "pytorch-CycleGAN-and-pix2pix," GitHub, updated 22 March 2024, accessed 4 January 2025, https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix/tree/master. In the bibliography a Google Colab notebook can be found containing all the used code in this research.

model were used. A batch size of 1 was used to accommodate the resource-intensive nature of the model, with training running for 200 epochs. The input images were preprocessed by resizing and random cropping. The training process was divided into two phases, a constant learning rate phase for the first 100 epochs, where the learning rate was fixed at 0.0002, and a linear decay phase over the last 100 epochs, during which the learning rate was gradually reduced to zero.

2.8. Evaluation Metrics

To evaluate the performance of the object detection models, several metrics were used that are widely accepted in the field of computer vision for evaluation and benchmarking. Before diving into these metrics, it is useful to first define four important concepts on which the metrics are based, namely Intersection over Union (IoU), true positives (TP), false positives (FP) and false negatives (FN). The IoU is the ratio of the intersection area to the union area of the predicted bounding box and the ground truth bounding box. It thus measures the overlap between the ground truth and predicted bounding boxes. When the IoU is 0, the predicted bounding box and the ground truth box do not overlap, when it is 1, the two boxes are identical.[109] A true positive occurs when the model correctly detects and localises an object, such as a tram or carriage, with a bounding box that sufficiently overlaps with the ground truth. The opposite is a false positive, which is when the model predicts an object that is not present in the picture or when the bounding box does not accurately correspond with the ground truth. Finally, a false negative happens when an object, like a carriage, is present in the picture but the model fails to detect it. True negatives (TN) are not relevant in an object detection context, since there are an infinite number of bounding boxes that should not be detected in an image.[110]

Based on these principles, two key evaluation metrics are calculated. Precision (P) is used to evaluate how precise the model is in its predictions, since this measures the proportion of correctly predicted objects (true positives) out of all predicted objects (true positives and false positives). ($Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$) Recall, on the other hand, measures the ability of the model to find all relevant

[109] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 10.
[110] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in *Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, ed. Anselmo C. Paiva et al. (Institute of Computing at Fluminense Federal University (IC-UFF), 2020), 238.

objects present in the picture, which is in other words the percentage of correct positive predictions among all given ground truths. ($Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$).[111]

Between precision and recall there is a trade-off, meaning that increasing the number of detected objects (higher recall) can result in more false positives (lower precision).[112] To evaluate how well a model balances this trade-off across different thresholds, the metric average precisions (AP) is used. AP gives a balanced assessment of precision and recall by considering the area under the precision-recall curve. When there is a high area under the curve, thus when precision stays high as the recall increases, the object detector is considered good.[113] The mean average precision (mAP) averages the AP values across all the classes or across multiple intersection over union thresholds. The closer the value of the AP and mAP to 1, the better the model.[114] In this study, the mAP50 and mAP50-95 are reported, respectively measuring the precision at an IoU of 0.5, and the mAP averaged over multiple IoU threshold ranging from 0.5 to 0.95. The mAP50 provides a more lenient assessment, counting a detection as correct if there is at least 50% overlap between the predicted and ground truth bounding boxes and the mAP50-95 is more comprehensive since it averages performance across different thresholds.

[111] Ibid.; Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 10; Hatir et al., "The deep learning method applied to the detection and mapping of stone deterioration in open-air sanctuaries of the Hittite period in Anatolia," 42-43.
[112] Terven, Córdova-Esparza, and Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," 1682-83.
[113] Siwar Bengamra et al., "A comprehensive survey on object detection in Visual Art: taxonomy and challenge," *Multimedia Tools and Applications* 83, no. 5 (2024): 14645-47, https://doi.org/10.1007/s11042-023-15968-9.
[114] Hatir et al., "The deep learning method applied to the detection and mapping of stone deterioration in open-air sanctuaries of the Hittite period in Anatolia," 43.

3. Setting the bar: benchmarking object detection models

As the field of machine learning is expanding at a rapid pace, with new object detection models being released quite often, it is necessary to first benchmark a few of these models and see their strengths and weaknesses.[115] By systematically comparing different models, it can be determined which model is the most effective for detecting objects in historical photographs. This chapter addresses the research question: "How do existing object detection models, such as YOLOv8, Faster R-CNN and RetinaNet, perform 'out of the box' on historical and modern photographs?" To answer this, a comparison of these three object detection models was performed. The methodological workflow of this chapter can be seen in Figure 4. The models were tested 'out of the box', meaning that they were tested using their pretrained weights without additional training on the project's datasets. Two different datasets were used, one containing 300 modern photographs and one made of 300 historical photographs. By using two different datasets, it was possible to detect if and how much performance of the existing models deteriorated on historical images. In total, four classes out of the 80 COCO classes were chosen to detect, specifically "person", "train", "boat", and "horse". These classes were chosen because they aligned with the objective of this thesis, namely developing an object detection model that can be used to detect sound sources on historical photographs in an urban context.

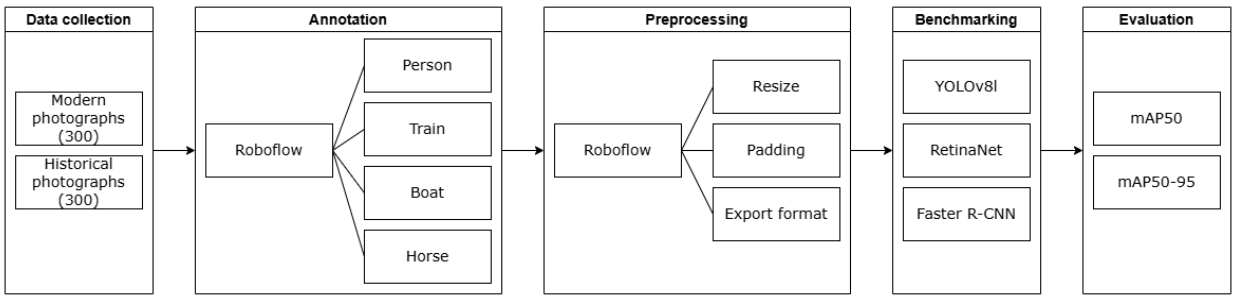


Figure 4: Workflow for the benchmarking of three object detection models.

The following sections will go into deeper detail on the performance on modern photographs, then on historical photographs and lastly these two will be compared. The evaluation will be done based on the mAP50 and the mAP50-95 metrics. Precision and recall will not be taken into consideration in this chapter, since the evaluation mode of both Faster R-CNN and RetinaNet do not provide these metrics by default. This evaluation forms a crucial foundation for the following two chapters, since the best performing model was selected for further training on custom datasets.

[115] Aditya Patra and Rae Crandall, "Machine Learning for Visually Impaired: Benchmarking Object Detection Models," *Journal of Student Research* 13, no. 2 (2024): 1, https://doi.org/10.47611/jsrhs.v13i2.6630.

3.1. Performance on modern photographs

Table 5 summarises the performance of each model on the modern dataset of 300 photographs from an urban context. As can be seen, YOLOv8l performed best overall, with a mAP50 of 0.755 and a mAP50-95 of 0.537. This result is even slightly higher than the official benchmark results of YOLOv8l on the COCO eval2017 dataset, a common dataset used to compare the performance of different object detection models, which can be seen in Table 6.[116] The second best performing model was Faster R-CNN, with a mAP50 of 0.476 and a mAP50-95 of 0.302, which is worse than the results on the COCO eval2017 dataset. RetinaNet had slightly worse results than Faster R-CNN, with a mAP50 of 0.428 and a mAP50-95 of 0.273.

Comparing the results per class, YOLOv8l outperformed the other two models in every category. The strongest performance of the model was on the "train" class, with a mAP50 of 0.937 and a mAP50-95 of 0.739, which is a very good result. Interestingly, both Faster R-CNN and RetinaNet failed to detect horses in the modern dataset, a peculiar phenomenon for which no direct explanation could be found. Despite the fact that the results slightly differed from the official mAP results on the COCO eval2017 dataset, the performance seems to reflect the fact that all three models were pretrained on datasets containing similar, present-day, images.

Metrics	Class	YOLOv8l	Faster R-CNN	RetinaNet
mAP50	Overall	0.755	0.476	0.428
	Person	0.561	0.518	0.385
	Train	0.937	0.787	0.823
	Boat	0.704	0.599	0.503
	Horse	0.816	0.00	0.000
mAP50-95	Overall	0.537	0.302	0.273
	Person	0.308	0.262	0.193
	Train	0.739	0.585	0.610
	Boat	0.446	0.361	0.291
	Horse	0.655	0.000	0.000

Table 5: Performance of the three object detection models on a dataset containing modern images.

[116] These numbers were taken from following sources: "Explore Ultralytics YOLOv8," Ultralytics, accessed 10 April 2025, https://docs.ultralytics.com/models/yolov8/; "Add FasterRCNN improved weights #5763," GitHub, accessed 10 April 2025, https://github.com/pytorch/vision/pull/5763; "Add RetinaNet improved weights #5756," GitHub, accessed 10 April 2025, https://github.com/pytorch/vision/pull/5756.

Model	mAP50	mAP50-95
YOLOv8l	0.695	0.529
Faster R-CNN	0.673	0.467
RetinaNet	0.618	0.415

Table 6: Performance of the three object detection models on the COCO eval2017 dataset.

3.2. Performance on historical photographs

When benchmarked on the 300 historical images, all models experienced a notable drop in performance, which can be seen in Table 7. All three models thus seemed to struggle with the additional visual challenges that historical photographs carry with them. Nevertheless, YOLOv8l still remained the top performer, with a mAP50 of 0.587 and mAP50-95 of 0.382 compared to a mAP50 of 0.315 and 0.253 for Faster R-CNN and RetinaNet and respectively a mAP50-95 of 0.186 and 0.147.

Looking at the different classes, it can be seen that the margin between YOLOv8l and Faster R-CNN narrowed for the "person" class. Although YOLOv8l still achieved the best results here, the difference with the modern dataset was smaller for Faster R-CNN, which could suggest that the two-stage detectors with a region-based approach like Faster R-CNN are more resilient to changes in visual style for well-represented, high frequency classes like person. In fact, it is interesting to see that for the "person" class in historical photographs, all of the models actually performed better than on their modern counterparts. A possible explanation for this could be that these photographs were "staged" more often, as it can be seen that people sometimes "posed" for the pictures, in ways that it would be clearer for the model to detect. The photographs containing horses still seemed to propose difficulties for Faster R-CNN and RetinaNet, although RetinaNet now did achieve some result, albeit it being pretty low.

Metrics	Class	YOLOv8l	Faster R-CNN	RetinaNet
mAP50	Overall	0.587	0.315	0.253
	Person	0.661	0.601	0.465
	Train	0.639	0.434	0.377
	Boat	0.402	0.224	0.158
	Horse	0.648	0.00	0.014
mAP50-95	Overall	0.382	0.186	0.147
	Person	0.383	0.324	0.242
	Train	0.481	0.307	0.271
	Boat	0.231	0.113	0.063
	Horse	0.450	0.000	0.010

Table 7: Performance of the three object detection models on a dataset containing historical images.

Figures 5 to 8 illustrate the results of the object detection models on the historical photographs and provide some useful insights.[117] The picture in Figure 5 is a relatively straightforward one, with the person and the two horses visible. As can be seen, YOLOv8l is the only model detecting both horses, even the one standing behind which is partially occluded. RetinaNet did detect the person in the picture, but it also drew many other random bounding boxes, a phenomenon which also occurs in the other figures. This is especially visible in Figure 6, with RetinaNet displaying many unnecessary bounding boxes. The other two models performed well on this photograph, despite it being a more difficult one due to the fact that there are many objects in the picture of which some are relatively small.

In Figure 7 and Figure 8, the "weak point" of YOLO can be observed. As Ali and Zhang state: "A major challenge is its difficulty in detecting small objects, particularly in cluttered or complex environments. Overlapping or occluded objects exacerbate this issue, as YOLO relies on anchor boxes and non-maximum suppression for object localization, which can lead to false negatives and overlooked detections."[118] As illustrated in Figure 7, YOLO was able to correctly identify the two horses, one of which is partially occluded, but here it is likely aided by the relatively large size and visual prominence of this horse. However, it failed to detect the two persons standing in the shadows in the background. Faster R-CNN successfully detected these figures, though it did not identify the horses. A similar pattern can be seen in Figure 8, with YOLO again detecting the horse, and even identifying the person partially hidden in the background. Nonetheless, it overlooked the coachman sitting atop the carriage, of whom only the upper body is clearly visible. Faster R-CNN again spotted this individual.

[117] For a bigger version of the original photographs, see annexes 1-4.
[118] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection."

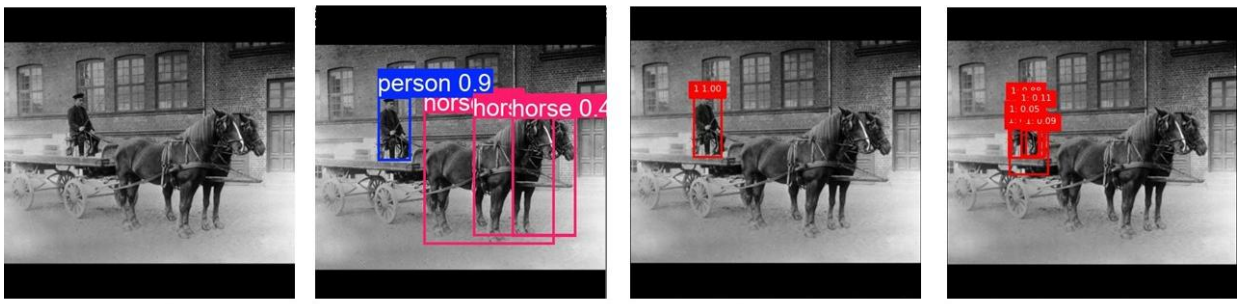


Figure 5: Comparison of the performance of the object detection models. (Source image: Hästvagn*. Photograph. Europeana (Malmö Museum). N.d. https://www.europeana.eu/nl/item/91672/MM_foto_612340.)*

Figure 6: Comparison of the performance of the object detection models. (Source image: Weltausstellung Paris 1867*. Photograph. Europeana (Albertina). 1867. https://www.europeana.eu/nl/item/15508/FotoGLV2000_22294.)*

Figure 7: Comparison of the performance of the object detection models. (Source image: Stadsgata*. Photograph. Europeana (Blekinge Museum). 1900-1915. https://www.europeana.eu/nl/item/916119/blm_item_68208.)*

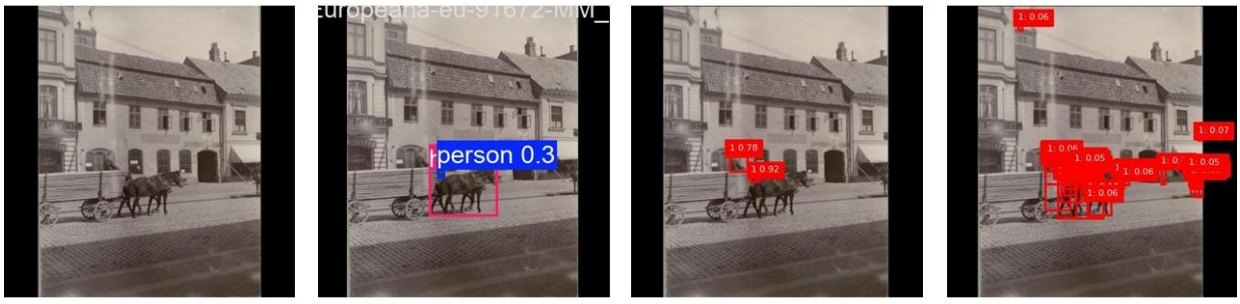


Figure 8: Comparison of the performance of the object detection models. (Source image: Ata, hästvagn*. Photograph. Europeana (Malmö Museum). 1910. https://www.europeana.eu/nl/item/91672/MM_foto_600289.)*

3.3. Comparative analysis

Comparing the results across the two datasets reveals important insights. In Table 8, the difference in performance between the modern dataset and the historical one can be seen, calculated by subtracting the results for the mAP50 and mAP50-95 of both datasets. Except for the "person" class and the "horse" class detected by RetinaNet, the domain gap between the data on which the models were pretrained and the historical dataset becomes clear. These results also underscore the limitations of using these pretrained object detection models 'out of the box' on historical data and thus the need of using transfer learning to apply the models in future research.

As stated before, Table 8 shows a clear overall performance advantage for YOLOv8l compared to Faster R-CNN and RetinaNet. It achieved the highest overall mAP, followed by Faster R-CNN and then RetinaNet. Based on these results, and the fact that it has a good balance of speed, accuracy, and community support, YOLOv8l was selected for further training.

Metrics	Class	YOLOv8l	Faster R-CNN	RetinaNet
ΔmAP50	Overall	0.168	0.161	0175
	Person	-0.100	-0.083	-0.080
	Train	0.298	0.353	0.446
	Boat	0.302	0.375	0.345
	Horse	0.168	0.000	-0.014
ΔmAP50-95	Overall	0.155	0.116	0.126
	Person	-0.075	-0.061	-0.049
	Train	0.258	0.278	0.339
	Boat	0.215	0.248	0.228
	Horse	0.205	0.000	-0.010

Table 8: The difference in performance on the modern and the historical dataset.

While the "train" class achieved the highest scores across models in modern images, it suffered a substantial performance drop compared to historical photographs. YOLOv8l achieved a mAP50-95 of 0.739 for this class on modern images, but on historical images the mAP50-95 was only 0.481, a difference of 0.258, the biggest difference of all the classes for this model. When looking at the results on the photographs, something interesting can be seen. The models seem to detect trams when they resemble their modern counterparts, but when there are more ancient types of trams, like horse-drawn trams, the model fails to detect those, pointing to the modern data that it was trained on. This can be illustrated by looking at Figure 9 to Figure 12.[119] Figure 9 and Figure 10 show more "modern" looking trams, which YOLO

[119] For a bigger version of the original photographs, see annexes 5-8.

was capable of detecting. Figure 11 and Figure 12 both display horse-drawn trams, of which YOLO only could detect the one in Figure 11. This could be due to the fact that this tram looks more like a modern one, while the one in Figure 12 is more different looking with its open sides.

This makes “train” an ideal class for further refinement in the next chapter: it is visually distinctive, relevant to soundscape analysis and prominent in many historical photographs, yet its performance in historical data leaves room for improvement. By selecting this class, the next chapter tested how well domain-specific fine-tuning can close the performance gap of YOLOv8l between modern and historical data.

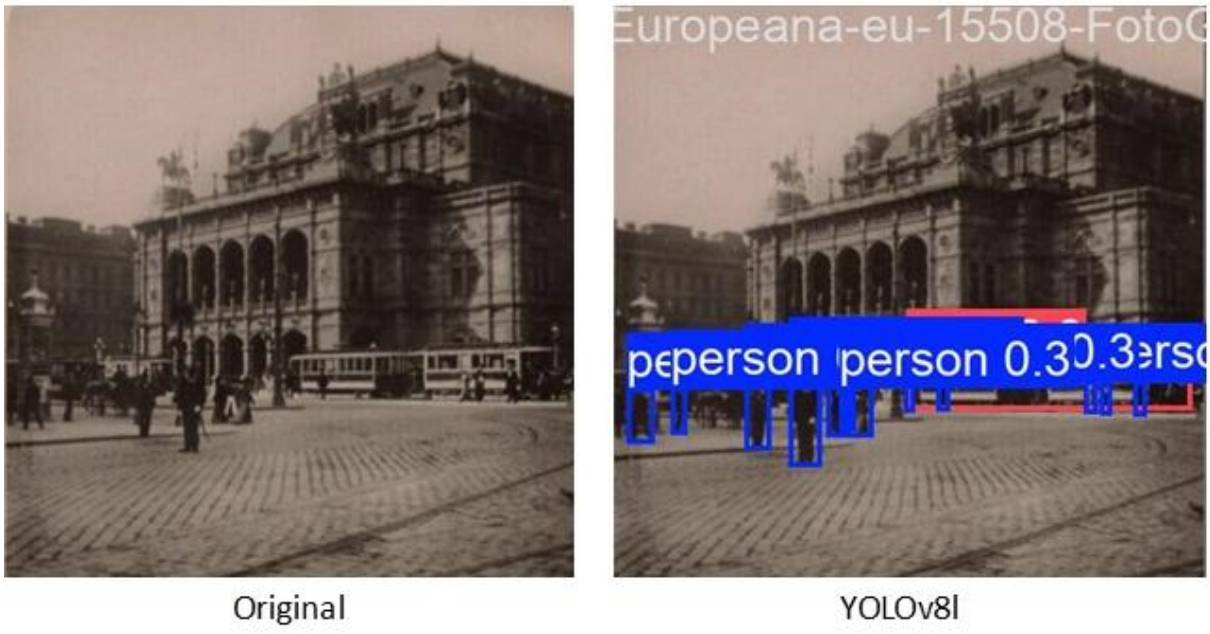


Figure 9: Performance of YOLO on historical photograph. (Source image: Wien Hofoper. *Photograph. Europeana (Albertina). 1908-1909. https://www.europeana.eu/nl/item/15508/FotoGLV2000_8382.)*

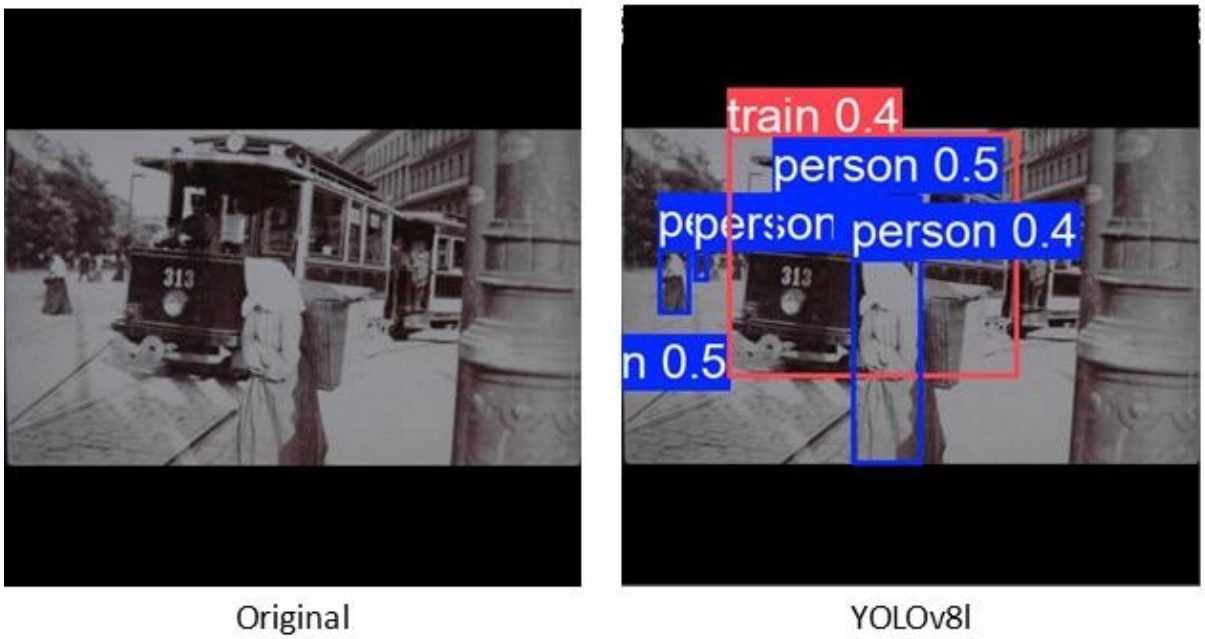


Figure 10: Performance of YOLO on historical photograph. (Source image: Wien, Szene mit Straßenbahn. *Photograph. Europeana (Albertina). 1906-1907. https://www.europeana.eu/nl/item/15508/Foto2005_177_32.)*

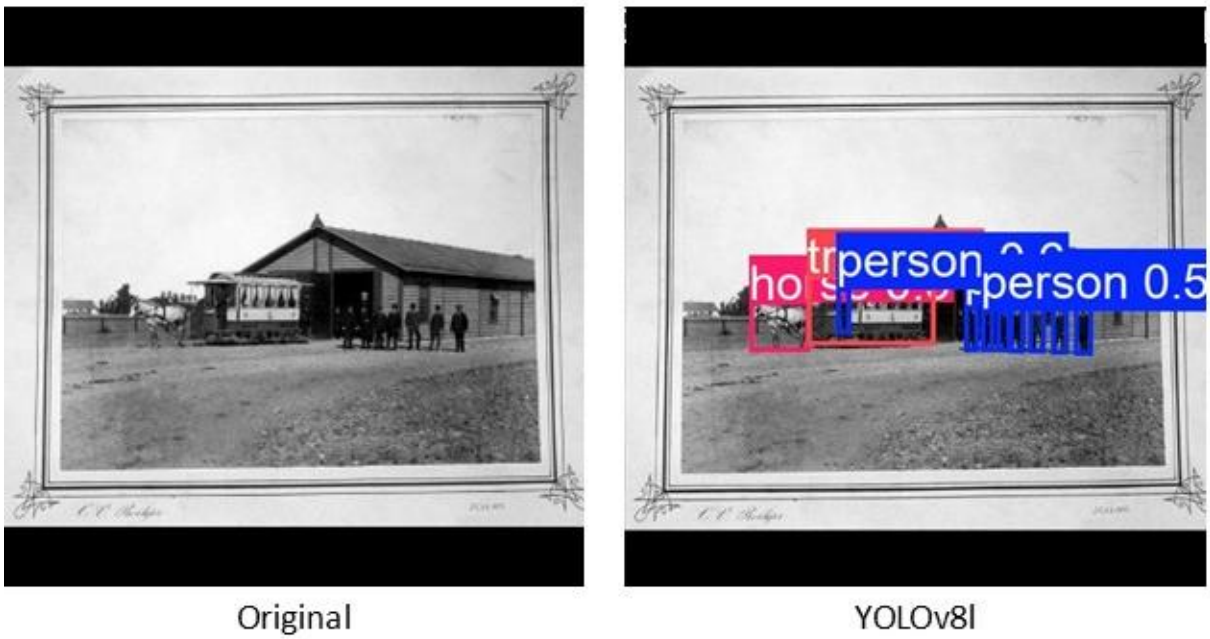


Figure 11: Performance of YOLO on historical photograph. (Source image: Spårvagn, hästspårvagn*. Photograph. Europeana (Malmö Museum). 1907. https://www.europeana.eu/nl/item/91672/MM_foto_611540.)*

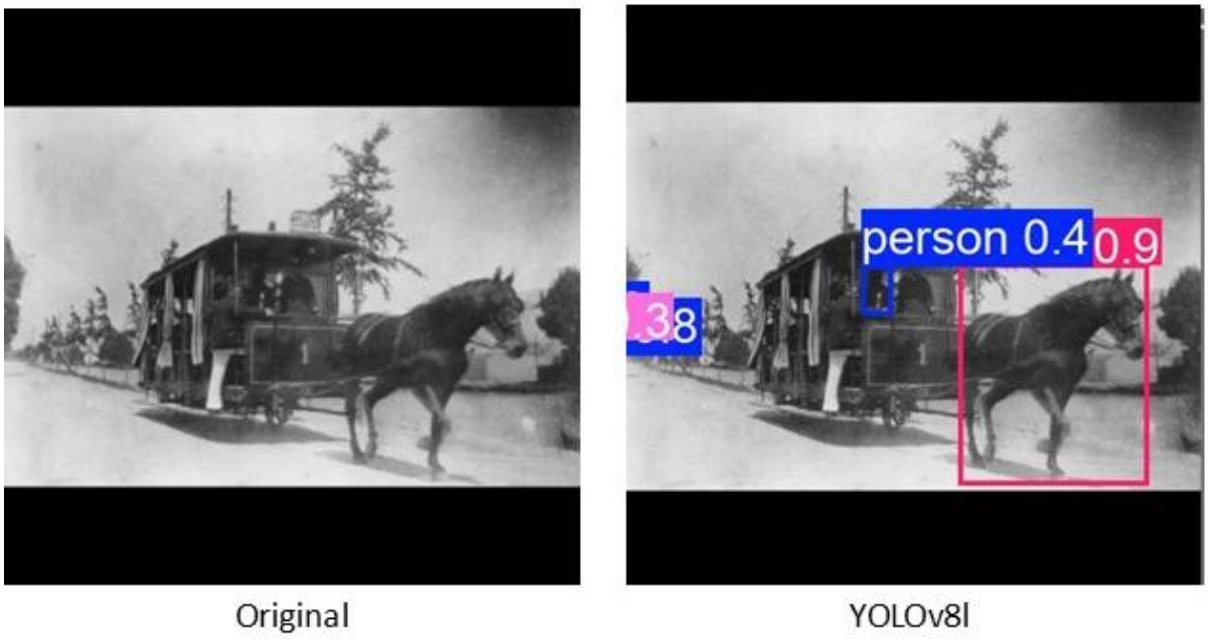


Figure 12: Performance of YOLO on historical photograph. (Source image: Gezicht op de Jutfaseweg te Utrecht met de paardentram Utrecht-Jutfaas/Vreeswijk. Photograph*. Europeana (Het Utrechts Archief). 1895. https://www.europeana.eu/nl/item/257/https___hetutrechtsarchief_nl_beeld_CBC12DD377C95F9E94E1FC6A5F2470FA.)*

4. Raising the bar: fine-tuning YOLOv8 on a pretrained class

Now that it has been established that out of the three object detection models, YOLOv8l performs best on historical photographs, this model is the most promising candidate for further development into one that can be used in soundscape research. Building upon the insights of the previous chapter, this chapter and the next one focused on applying transfer learning with the objective to fine-tune a model trained on modern data to historical data. In this chapter, one of the 80 COCO classes that the YOLOv8l model was pretrained on, more specifically the class “train”, was selected for further experimentation to answer the research question: “How does training an existing class in the model, such as trains/trams, using historical images affect the accuracy of the model?”. Since fine-tuning leverages the weights and parameters of a pretrained model to perform a related task, it was first applied to a class that the model had already encountered in its original training, with the rationale being that the model might need less data for this.[120]

The model was trained on eight different historical datasets, one base dataset containing a total of 422 pictures of which 295 were used for training, and then six datasets where these training images were augmented using different techniques. These ranged from single augmentations, namely brightness, exposure, blur, saturation and noise, to combined augmentations. Training was done on Google Colab, using a T4 GPU for 100 epochs. The full workflow is visualised in Figure 13.

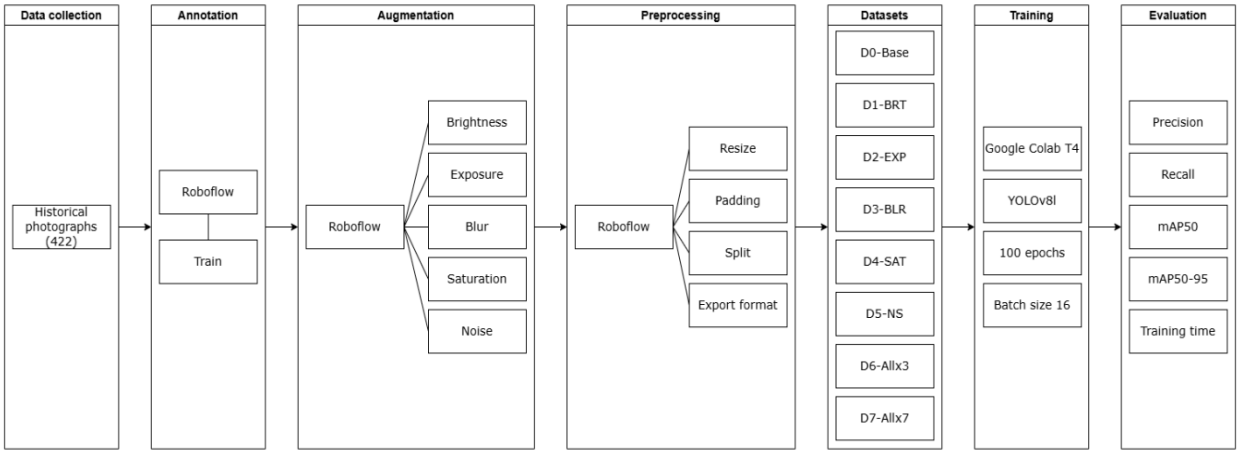


Figure 13: Workflow for training YOLOv8l on a "known" class, "train".

[120] Kulkarni et al., "Short Classification of Cultural Heritage Sites Using Transfer Learning." 3.

The results of the training experiments can be found in Table 9.[121] Before going into detail, it can be useful to quickly revisit the benchmark values established in the previous chapter, which can be seen in the first two rows of the table. Untrained on the new data, YOLOv8l achieved a very high precision of 0.951 and a recall of 0.785 on the "train" class in the modern dataset. The mean average precision scores were also good, reaching 0.937 for mAP50 and 0.739 for mAP50-95. When looking at the results of the 'out of the box' model on the "train" class with historical photographs, a drop in performance could be seen. Precision was 0.760, recall 0.465, mAP50 0.639 and mAP50-95 0.481. These differences highlighted the visual gap between the modern images on which the model was previously trained and the historical photographs on which the model has to perform for this research.

The first experiment trained the YOLOv8l model on the unaugmented dataset, D0-Base, and served as a reference point for evaluating the impact of the different augmentation techniques. Being the smallest dataset, it took the shortest time to train, with a training time of one hour. It can be seen that this model achieved a precision of 0.714, recall of 0.531, mAP50 of 0.583 and a mAP50-95 of 0.395. When this is compared with the benchmark results, a drop in performance can be noticed. The model thus did not perform better after retraining it with the smallest dataset of 295 training images. The only metric that improved was the recall, meaning that the model did identify more trains or trams that it previously did not detect, but at the same time the precision went down, signifying that it also predicted more false positives, so detecting trains that are in fact not trains. In other words, the model detected more objects overall but the predictions were less accurate.

The question then arose whether the use of augmentation could enhance the performance of the model. As can be observed in Table 9, compared to D0-Base, the mAP50 and mAP50-95 increased for all the different augmentation techniques. In fact, precision and recall also increased for each technique, except for the model trained on a dataset containing images where the exposure was changed, where the precision slightly dropped. The augmentations that had the biggest impact on the performance of the model were brightness, saturation and noise. D1-BRT reached a precision of 0.822 and recall of 0.559, a mAP50 of 0.669 and a mAP50-95 of 0.488, which was the highest mAP50-95 of all the different models. This suggests that this model outperforms the other models in predicting bounding boxes that are close to the ground truth boxes. The model trained on D4-SAT had the highest precision, with a value of 0.840. This means that out of all the detections the model made, 84% were true positives, signifying the fact that the model is fairly reliable when it predicts an object as a train. However, as often stated, there is a trade-off between precision and recall, which can be illustrated by the fact that D4-SAT had the second lowest recall of the different augmentation techniques. Finally, D5-NS had the highest recall and the highest

[121] For the visualisations of the model performance, see annex 9-16.

mAP50. The recall of 0.599 indicates that the model correctly identified approximately 60% of all the actual objects present in the photographs, or in other words it missed around 40% of the trams. The models trained on the datasets with exposure and blur performed slightly worse when comparing the mAP50-95 values, respectively 0.459 and 0.461. The precision of D2-EXP was the lowest of all, with a score of 0.701 indicating that 70% of the positive detections were correct, which is 14% lower than the model using the D4-SAT dataset. However, it achieved the second highest recall, with a value of 0.588.

The final two models, D6-Allx3 and D7-Allx7 used a combination of all these augmentation techniques, respectively augmented three and seven times. Just like the “single” augmentations, the training data for D6-Allx3 was also augmented three times, but it just incorporated a mix of all the different augmentations. D6-Allx3 achieved the highest precision of all the different models, with a score of 0.871 indicating that 87% of the predictions were true positives. With a value of 0.675 the model also had the highest mAP50 of all the different models, but the mAP50-95 of 0.464 was a bit lower than some of the others. These results suggest that a combination of augmentation techniques, which are all based on factors that could be present in historical photographs, can improve model performance, particularly in terms of precision and overall object detection, as illustrated by the mAP50. The mAP50-95 was only average however, which could imply that while it detected objects correctly, the bounding box localisation was not the most precise at higher IoU thresholds.

What is interesting to see, is the fact that increasing the augmentation factor even further, from three to seven, did not lead to any drastic improvements. The precision, recall and mAP50 of D7-Allx7 were slightly lower than the scores for D6-Allx3. Only for the mAP50-95 did augmenting the training data seven times have a positive effect, with a score of 0.485 compared to 0.464 in D6-Allx3. The model does seem to be better at detecting objects across a wider range of IoU thresholds. The gains were marginal however, suggesting that the law of diminishing returns applies here, where more augmentation does not necessarily lead to model generalisation.[122] Data augmentation is thus a useful tool, but it cannot fully replace collecting more, new, data.

[122] Eder Arley Leon-Gomez, Andrés Marino Álvarez-Meza, and German Castellanos-Dominguez, "Cross-Dataset Data Augmentation Using UMAP for Deep Learning-Based Wind Speed Prediction," *Computers* 14, no. 4 (2025): 136, https://doi.org/10.3390/computers14040123.

Name	Augmentation	Precision	Recall	mAP50	mAP50-95	Training Time (h)
Benchmark Modern (Class Train)	/	0.951	0.785	0.937	0.739	/
Benchmark Historical (Class Train)	/	0.760	0.465	0.639	0.481	/
D0-Base	No augmentation	0.714	0.531	0.583	0.395	0.592
D1-BRT	Brightness	0.822	0.559	0.669	0.488	1.399
D2-EXP	Exposure	0.701	0.588	0.655	0.459	1.470
D3-BLR	Blur	0.790	0.565	0.645	0.461	1.359
D4-SAT	Saturation	0.840	0.563	0.656	0.478	1.299
D5-NS	Noise	0.805	0.599	0.674	0.468	1.352
D6-Allx3	All of the above , x3	0.871	0.574	0.675	0.464	1.425
D7-Allx7	All of the above, x7	0.861	0.559	0.674	0.485	3.129

Table 9: Performance results of YOLOv8l trained on different datasets.

Despite the various improvements the augmented models made with regards to the model trained on the base dataset, it is important to note that none of the models managed to fully bridge the gap with the performance of YOLOv8l on the modern dataset. When compared to the benchmark on the historical dataset, it can be seen that an improvement of 10% was made on both precision and recall. As a result, the model became significantly better at identifying relevant objects and avoiding false positives. Looking at the mAP50 and mAP50-95 scores however, only marginal improvement can be seen. This suggests that while the model was better at classifying objects, the precision of the bounding box localisation did not advance to the same extent.

Figures 14-20 visually represent the performance of the different models, each showing how the eight models performed on the same historical images.[123] Figure 14 and Figure 15 demonstrate the improvement of various models compared to D0-Base. In Figure 14, most models correctly identified the tram in the centre of the photograph, which D0-Base did not capture, albeit it being with different confidence scores which can be seen next to the class name. D4-SAT and D6-Allx3 also failed to detect

[123] For a bigger version of the original photographs, see annexes 17-23.

this. Figure 15 illustrates that, except for D4-SAT again, all the models detected the tram. D7-Allx7 detected an additional object at the right of the images, which turned out to be a false positive.

Figure 16, Figure 17 and Figure 18 show improvements in some of the models in reducing false positives, meaning that the models correctly stopped detecting objects that were not trams compared to D0-Base. Figure 16 clearly demonstrates this, with D0-Base classifying the building at the left of the photograph as a tram and all the augmented models correctly detecting only the tram. In Figure 17, D2-EXP, D4-SAT and D5-NS still misclassified a column as a tram, while all the others improved. This figure also reveals that no model was able to capture the three trams overlapping in the foreground as three separate objects. Figure 18 shows that, just like D0-Base, D1-Brightness misclassifies a carriage as a tram, while all the other models improved. D4-SAT however introduced a new false detection in the background.

Finally, Figure 19 and Figure 20 show interesting cases where the models trained on the augmented data did not always improve compared to D0-Base and where they performed inconsistently. The photograph in Figure 19 shows three trams in total, one at the foreground, which is pretty obvious to detect, than a tram in the middle which is slightly obscured by a carriage and horse in front of it, and finally a third tram in the background. D0-Base was able to detect the largest tram and the second one that was obscured by the carriage, while some of the augmented models failed to detect it. Only D2-EXP, D4-SAT and D6-Allx3 were capable of detecting it, albeit it with slightly oversized bounding boxes that also took in a part of the carriage as a tram. Interestingly, D3-BLR was the only model that detected the smallest tram in the background. Figure 20 depicts a half-obscured tram, that only D0-Base and D2-EXP were able to detect, all the other models missed it.

Since the combination of the different augmentation strategies gave balanced results, this combined approach was selected to augment the training data for the detection of a new class, “carriage”, in the next chapter.

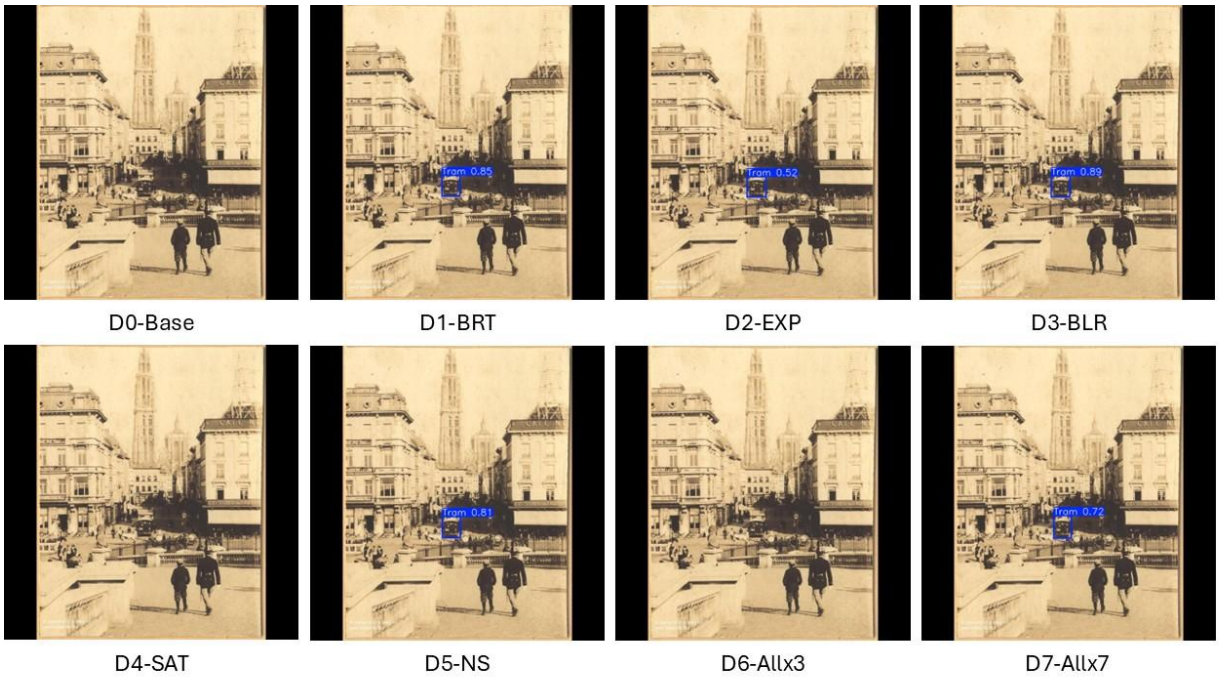


Figure 14: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: Suikerrui, vanaf het Zuiderterras, Antwerpen*. Photograph. FelixArchief. 1909. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_7087&dtnr=1224_40&dtrecordid=32372&page=1&pageSize=10&type=copy.)*

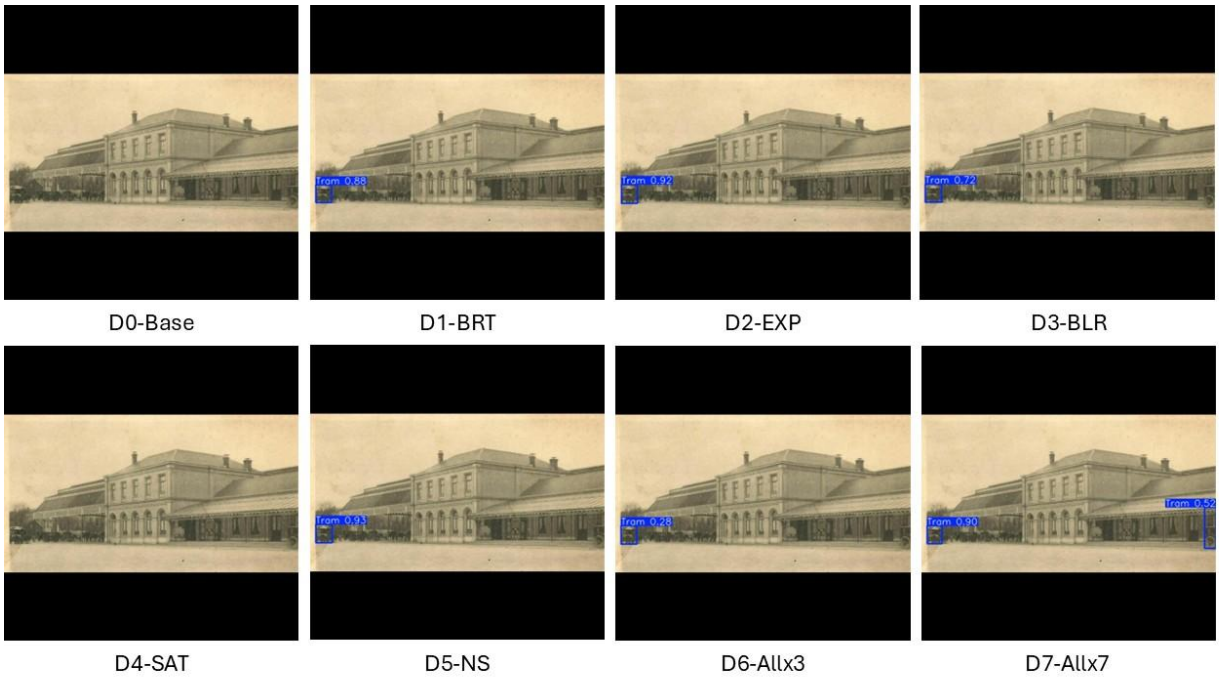


Figure 15: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: Gezicht op het S.S.-station Den Haag S.S. te Den Haag. *Photograph. Europeana (Het Utrechts Archief). 1899. https://www.europeana.eu/nl/item/257/https___hetutrechtsarchief_nl_beeld_025EB5A6A32053F7A8961CD4B2798CA9.)*

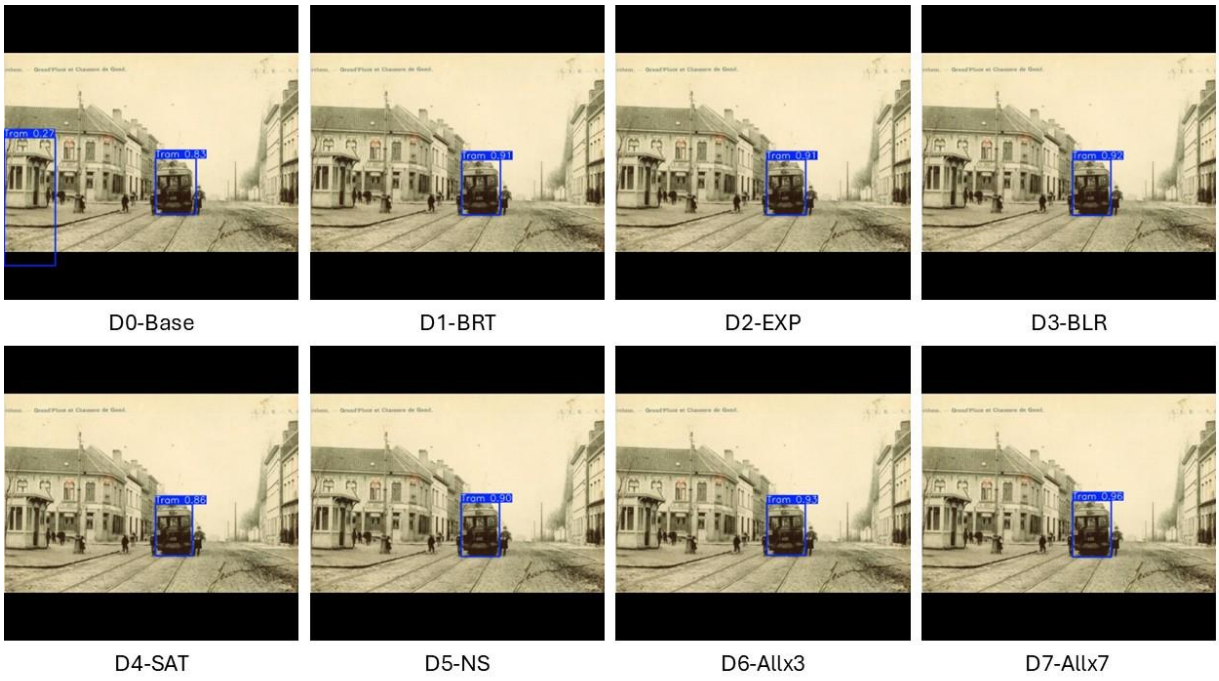


Figure 16: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: Sint-Agatha-Berchem: Grand'Place en Gentsesteenweg met tram*. Photographs. Erfgoedbank Brussel. N.d. https://erfgoedbankbrussel.be/mediabank/detail/f8619c70-3828-80b1-d9bb-17fc6a401c34/media/4876b394-3fe3-fb1c-c95b-618d29dd33ff?mode=detail&view=horizontal&q=Berchem%20&rows=1&page=59&fq%5B%5D=search_s_entity_name:%22Objecten%22.)*

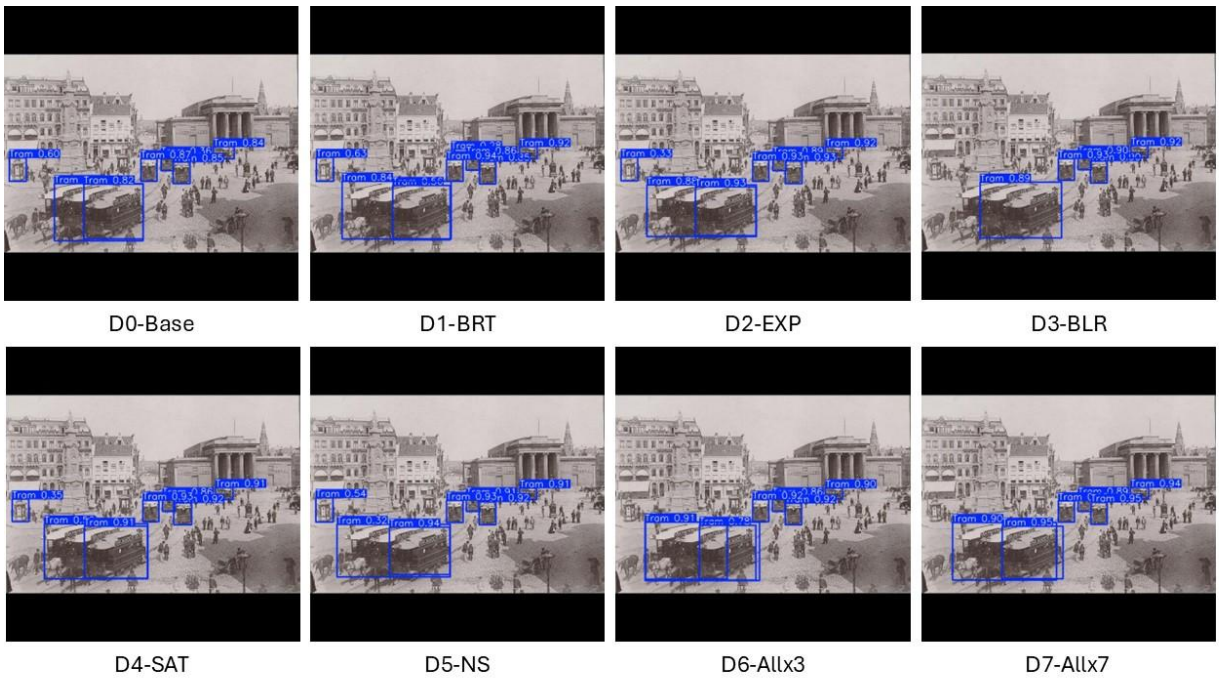


Figure 17: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: Amsterdam. *Photograph. Europeana (Rijksmuseum). 1880-1920. https://www.europeana.eu/nl/item/90402/RP_F_F19497.)*

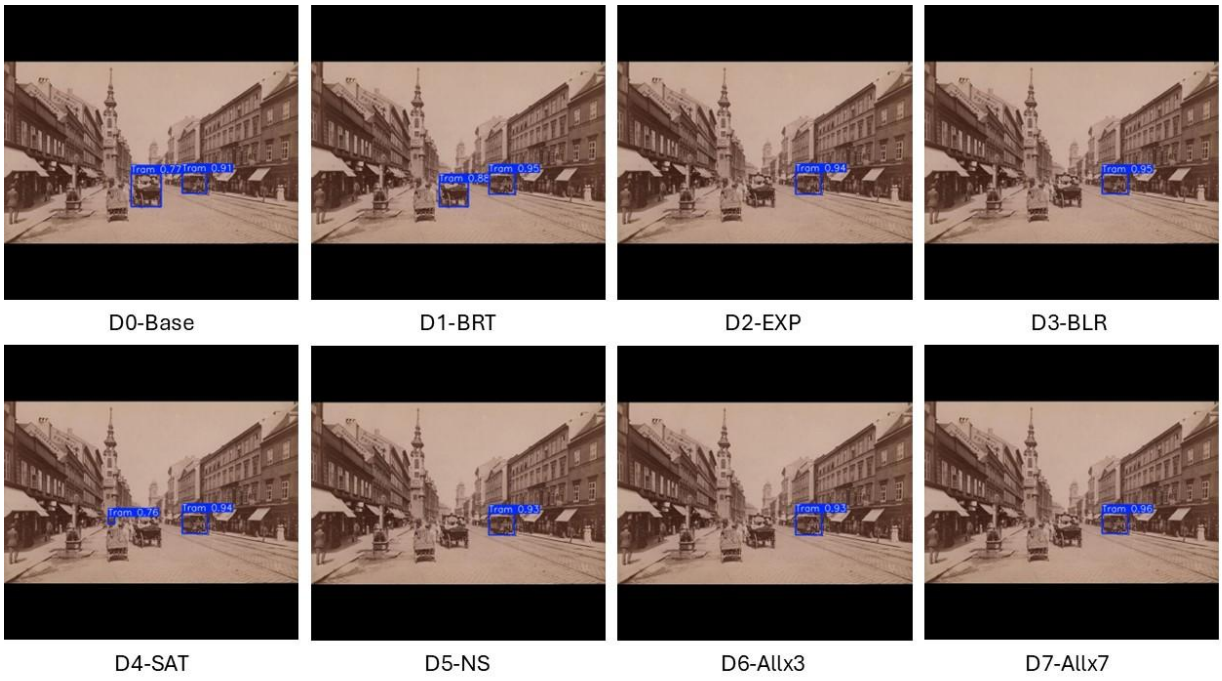


Figure 18: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: Wien, Mariahilfer Straße, Blick zur *Stadt. Photograph. Europeana (Albertina). C. 1890.* *https://www.europeana.eu/nl/item/15508/Foto2005_166_77.)*

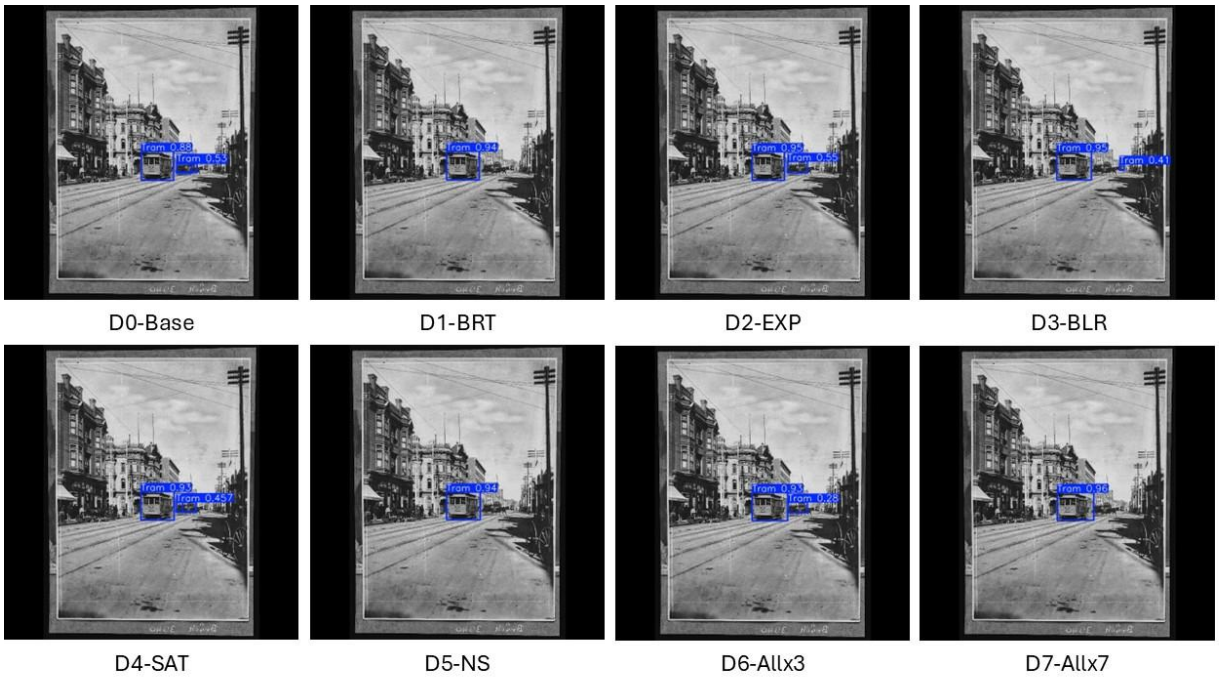


Figure 19: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: San Diego. 5th st., San Diego, Cal*. Photograph. Europeana (Deutsche Fotothek). 1903.* *https://www.europeana.eu/nl/item/463/item_T6IBXVO34VDXMPWACLDHKWT62IQUFB62.)*

Figure 20: Comparison of results of YOLOv8 trained on different datasets containing trains. (Source image: Korte Nieuwstraat, hoek Melkmarkt, Antwerpen. *Photograph. FelixArchief. 1931. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_6678&dtnr=1224_40&dtrecordid=31973&page=1&pageSize=10&type=copy.)*

5. Beyond the known: transfer learning for “unknown” historical objects

The previous chapter illustrated that moderate performance gains can be made by fine-tuning an existing object detection model with a relatively small dataset on one of the 80 classes that the model already was trained on before. But often, the existing classes that the model was trained on do not fully align with the research objective for which the model will be used.[124] For instance, in the case of this research into soundscapes, there are many objects one can think of that produce sound and appear in historical photographs that are not part of the 80 COCO classes. In that case transfer learning can still be used to add a new class to the model, even if it has no prior knowledge of this class. This chapter explored the performance of YOLOv8l when trained to detect a new class, “carriage”, demonstrating how transfer learning can be applied in further research by adding more classes. Central in this chapter is the question: “How can an object detection model be extended to include a new class, such as carriages, and how well does the model perform on this new class?”.

For the training of the model, two different historical datasets were used. The first one was a base dataset, containing a total of 1159 pictures, of which 821 were used for training the model. Next, this dataset was augmented by using a combination of the different augmentation techniques mentioned in the previous chapter. The augmentations were applied seven times, resulting in a dataset of 6085 pictures, of which 5747 were training data. This final dataset was substantially bigger than the datasets used in the previous chapter, because it was expected that the model would require more training data to learn this new class. The training was then done via Google Colab, using a T4 GPU for 90 epochs. The full workflow is visualised in Figure 21.

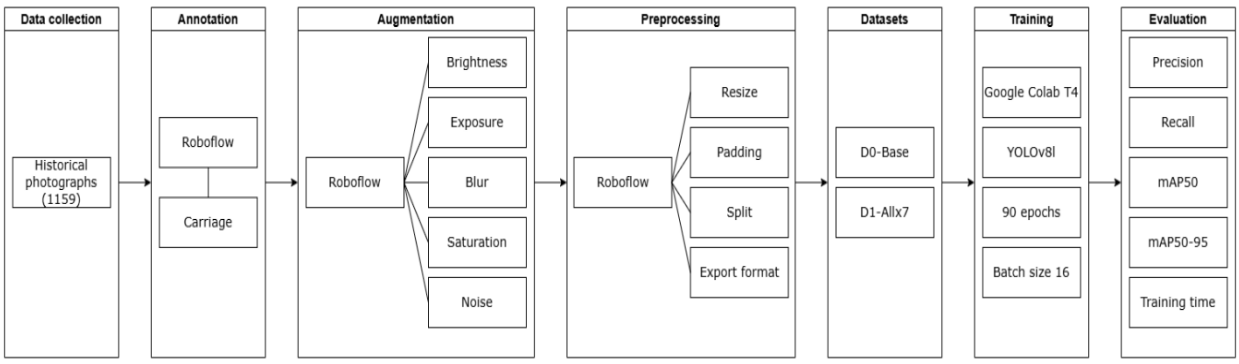


Figure 21: Workflow for training YOLOv8l on an "unknown" class, "carriage".

[124] Wevers, "Scene Detection in De Boer Historical Photo Collection," 601.

The results can be found in Table 10.[125] The base model reached a precision of 0.605, a recall of 0.447, a mAP50 of 0.461 and a mAP50-95 of 0.251, with a training time of one hour and 47 minutes. Comparing this with the results of model D6-Allx3 of the previous chapter (0.871 precision, 0.574 recall, 0.675 map50, 0.464 mAp50-95), which was trained on a known class using a similar number of training images, it becomes clear that the model struggled more with this new class. The performance gap strongly points to the fact that training a model on a new class thus requires a larger dataset to achieve comparable results as training one on a known class.

To address this performance gap, a second model, D1-Allx7, was trained on an augmented version of the base dataset. Due to the large number of pictures, training this model took considerably more time, with a training time of more than seven hours and a half. The model improved however, with a precision of 0.763, recall of 0.452, mAP50 of 0.577 and mAP50-95 of 0.326. While the recall only improved marginally, the precision improved with nearly 16%, the mAP50 with nearly 10% and the mAP50-95 with around 7%. However, the gap with the "known" class can still be seen, and the model is still quite far from an ideal performance. The recall of 0.452 means that the model detected only 45.2% of the total carriages in the photographs, which is not enough for data collection in soundscape research.

Name	Augmentation	Precision	Recall	mAP50	mAP50-95	Training Time (h)
D0-Base	No augmentation	0.605	0.447	0.461	0.251	1.466
D1-Allx7	All discussed augmentations, x7	0.763	0.452	0.557	0.326	7.325

Table 10: Results of the different models.

To visualise the differences between the base model, D0-Base, and the augmented model, D1-Allx7, a few sample outputs were included, which can be seen in Figure 22 to Figure 32.[126] These figures show a comparison between the two models, with their performance on the same photographs. Figure 22 and Figure 23 show a clear improvement in the ability of D1-Allx7 to differentiate between carriages and trams, with which D0-Base struggled. The performance of D1-Allx7 on Figure 22 is further also impressive because it managed to detect a partially visible carriage on the right of the photograph. As stated before, YOLO models often struggle with small and occluded objects, so it is interesting to see that in this picture the model succeeded to detect the carriage.[127] Figure 24 reinforces the conclusion that D1-Allx7 is better in discerning between carriages and trams, as it can be seen here that it no longer classified the horse-

[125] For the visualisations of the model performance, see annex 24-25.
[126] For a bigger version of the original photographs, see annexes 26-36.
[127] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," 27.

drawn tram as a carriage. Furthermore, while D0-Base only detected one of the three carriages in the background, D1-Allx7 detected all three, pointing towards the improvement in sensitivity to smaller and more distant targets. In Figure 25, the ability of D1-Allx7 to detect smaller objects in the background becomes clear again, with the rather small carriage at the very back of the photograph being noticed by the model. Figure 26 demonstrates D1-Allx7's enhanced ability to handle occlusion, since it correctly distinguished the two carriages that are very close together, which D0-Base merged into a single object.

The results are not uniformly positive, however. The photograph in Figure 27 depicts both carriages and old cars. D0-Base mislabelled all cars as carriages, whereas D1-Allx7 showed improvement by correctly ignoring many of the cars. It still misclassified one vehicle however, albeit with a lower confidence score, which could point toward progress. In Figure 28, D1-Allx7 improved by avoiding a false detection on a building façade that D0-Base wrongly labelled as a carriage. The augmented model was also capable of detecting one additional carriage that the base model missed. However, several carriages, which are depicted rather clear, still go undetected by both models. In Figure 29, D0-Base outperformed D1-Allx7, detecting one carriage on the left of the picture, which D1-Allx7 did not see. D1-Allx7 also produced a false positive by detecting an unrelated object on the right side of the image. Finally, Figure 30, Figure 31 and Figure 32 are cases where the models still struggle. Figure 30 shows a complex scene with many occluded carriages. Although D0-Base detects one more carriage here than D1-Allx7, neither of the models performed particularly well. In Figure 31, a very obvious carriage is missed entirely by both models. Figure 32 illustrates the difficulties with detecting the carriage in the shadow at the left of the images, which both models did not see.

Figure 22: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Laken: Maria-Christinastraat. *Photograph. Erfgoedbank Brussel. 1930. https://erfgoedbankbrussel.be/mediabank/detail/38cb5de4-b70e-56ab-0a59-4587cc60214e/media/cde45153-8f39-f99b-e7e2-d54274a4aaa3?mode=detail&view=horizontal&q=laeken%20tram&rows=1&page=.)*

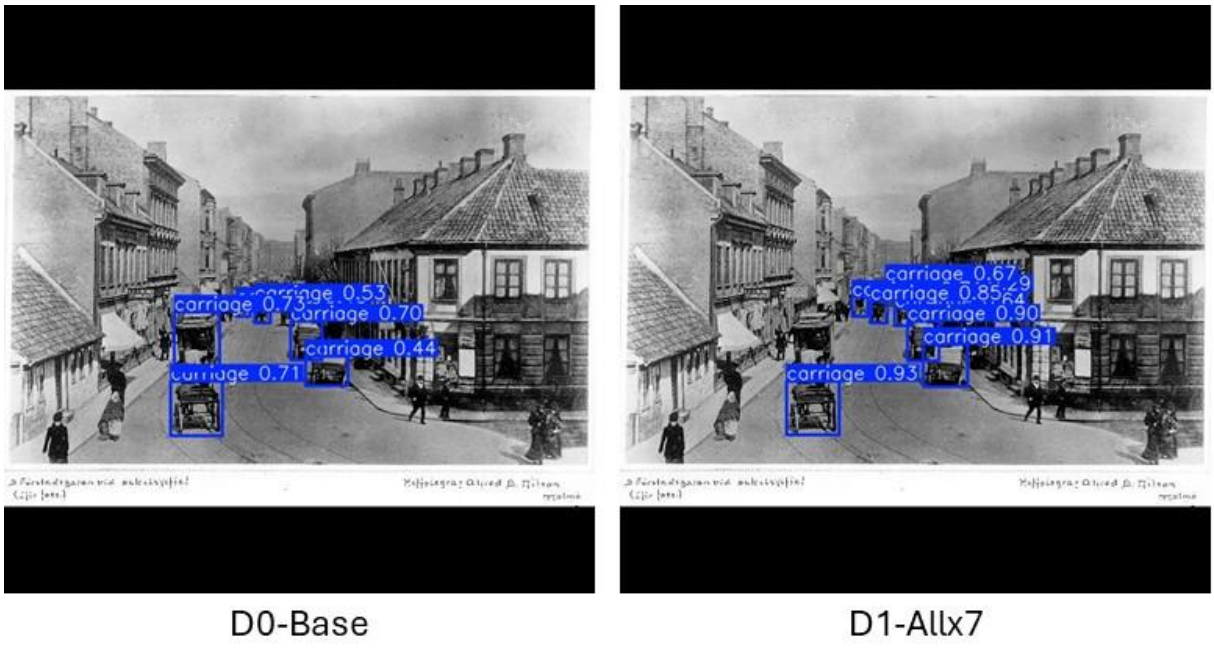


Figure 23: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Spårvagn, hästvagn. *Photograph. Europeana (Malmö Museum). 1895-1905. https://www.europeana.eu/nl/item/91672/MM_foto_612814.)*

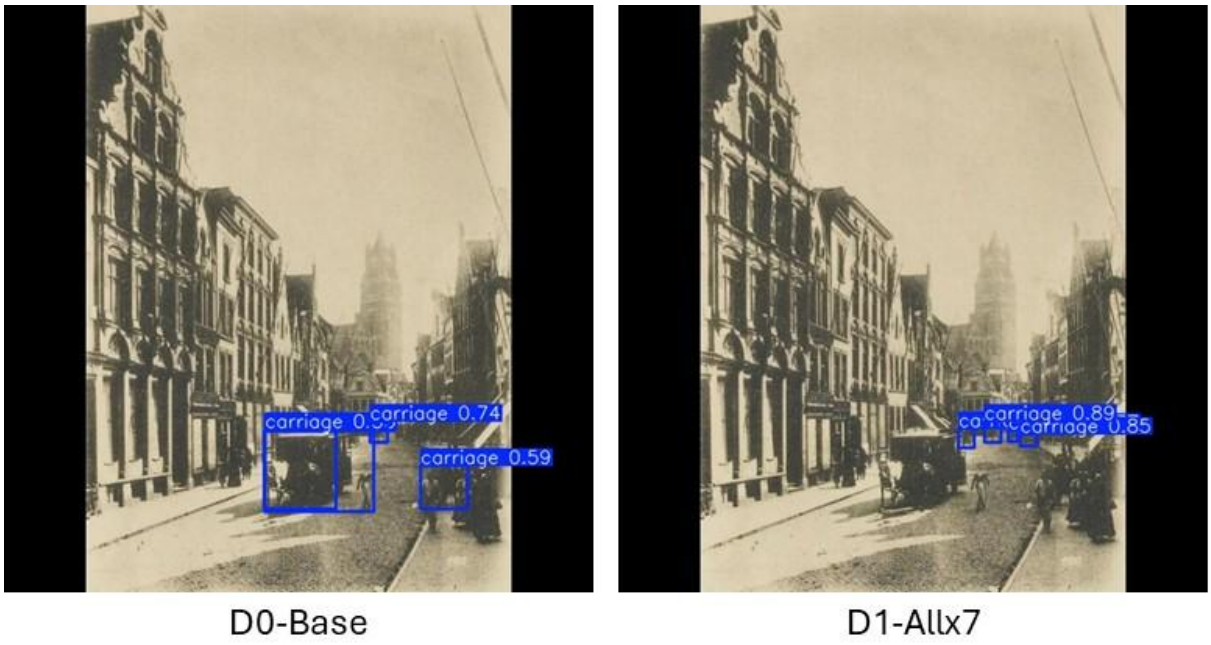


Figure 24: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Zicht op de Steenstraat tussen de Sint-Niklaasstraat en het Simon Stevinplein*. Photograph. Europeana (Stadsarchief Brugge). 1894. https://www.europeana.eu/nl/item/534/377edd52_8121_4ac2_8070_e484627e8a45.)*

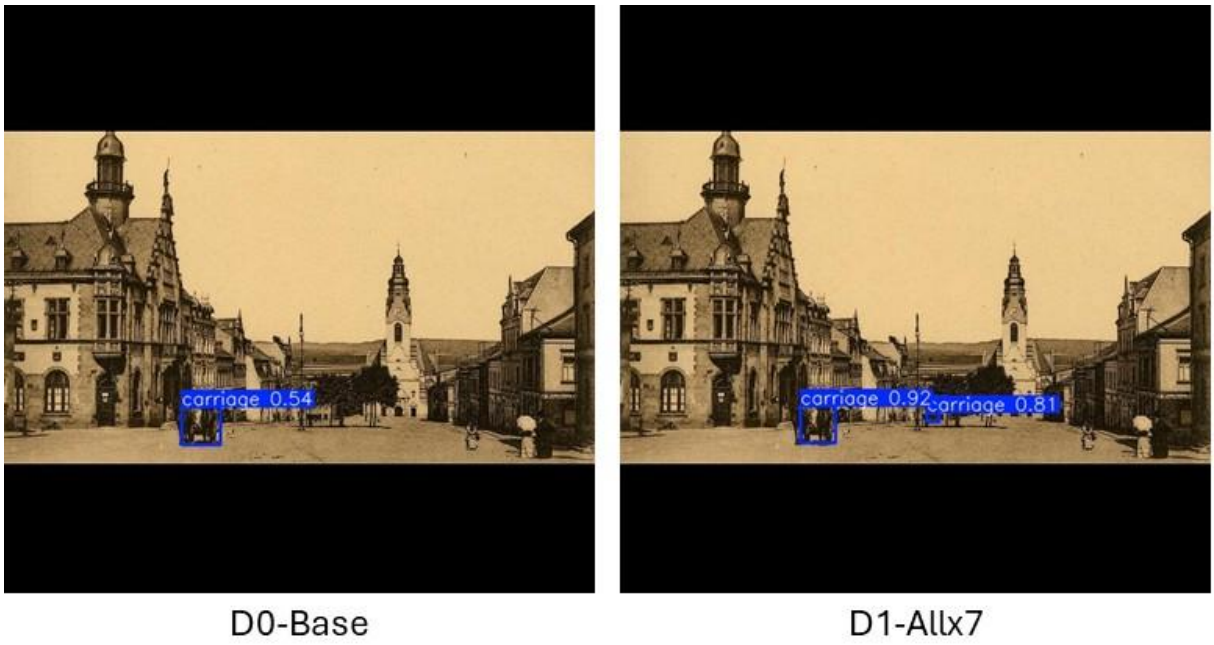


Figure 25: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Adorf. Markt. *Photograph. Europeana (Deutsche Fotothek). 1907. https://www.europeana.eu/nl/item/437/item_OKFKRYYUJPVTLRFZR6IUKJHTEJQJ6DS5.)*

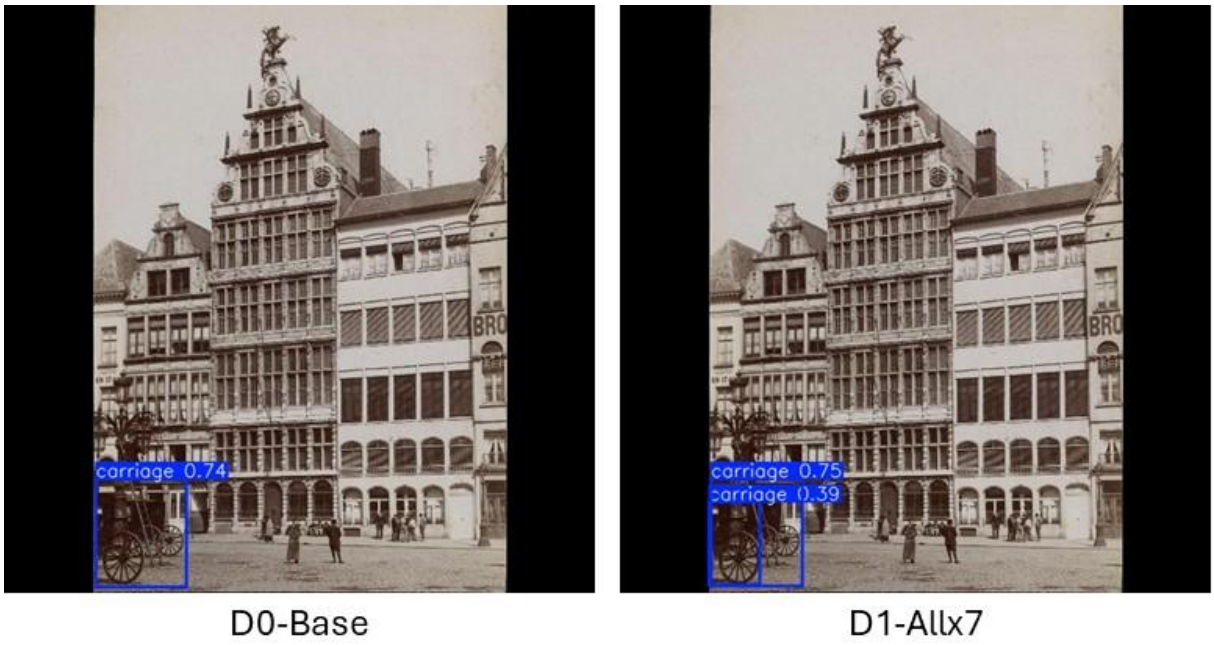


Figure 26: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Sint Joris Gildehuis in Antwerpen*. Photograph. Europeana (Rijksmuseum). 1880-1920. https://www.europeana.eu/nl/item/90402/RP_F_F16285.)*

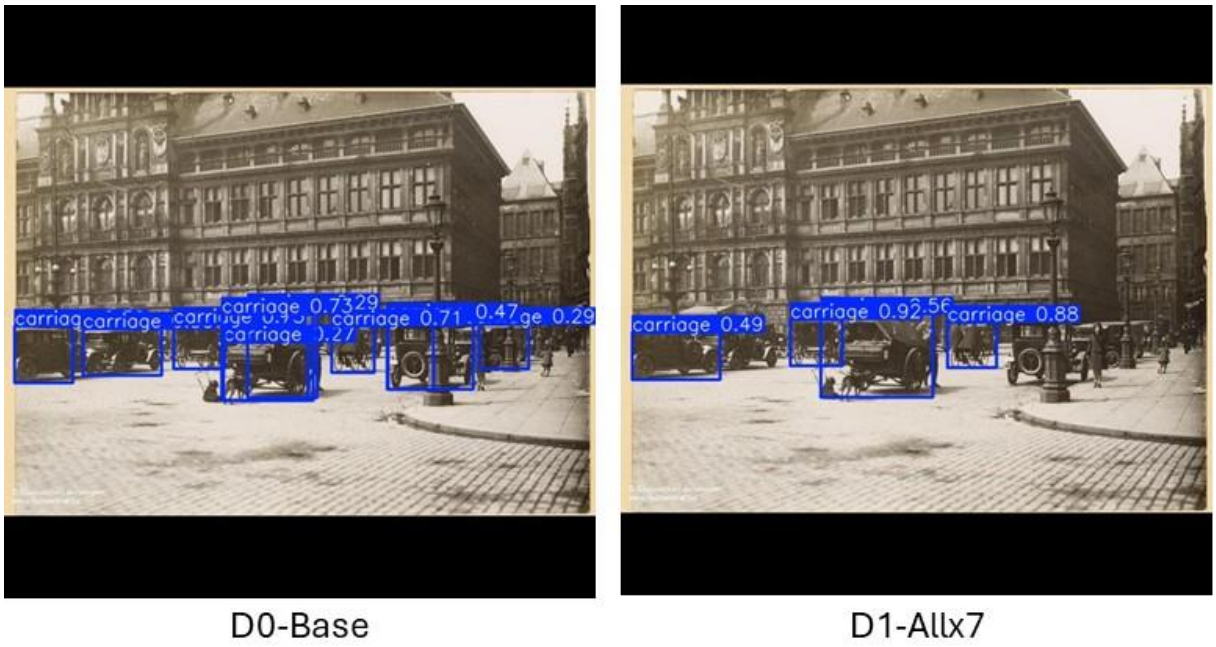


Figure 27: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Grote Markt: de linkervleugel van het Stadhuis, Antwerpen 1928. *Photograph. FelixArchief. 1928. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_3479&dtnr=1224_40&dtrecordid=28872&page=1&pageSize=10&type=copy.)*

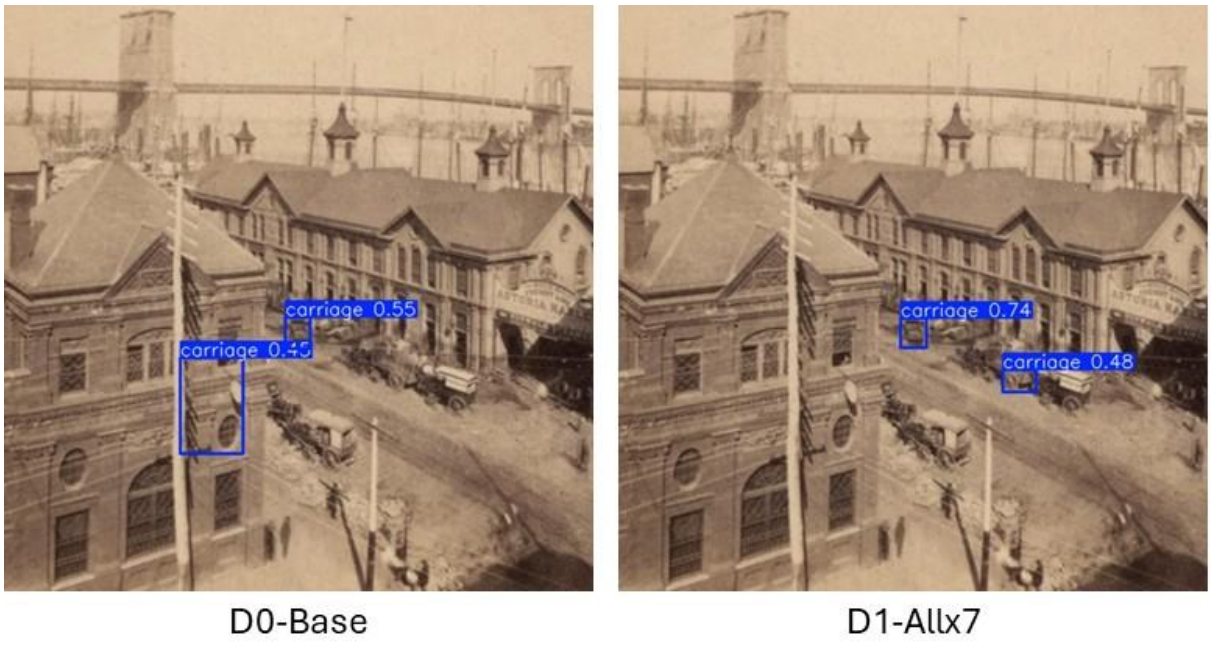


Figure 28: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: East River Bridge N.Y.*. Photograph. Europeana (Swedish National Museum of Science and Technology). 1860-1880. https://www.europeana.eu/nl/item/916118/S_TEK_photo_TEKA0113700.)*

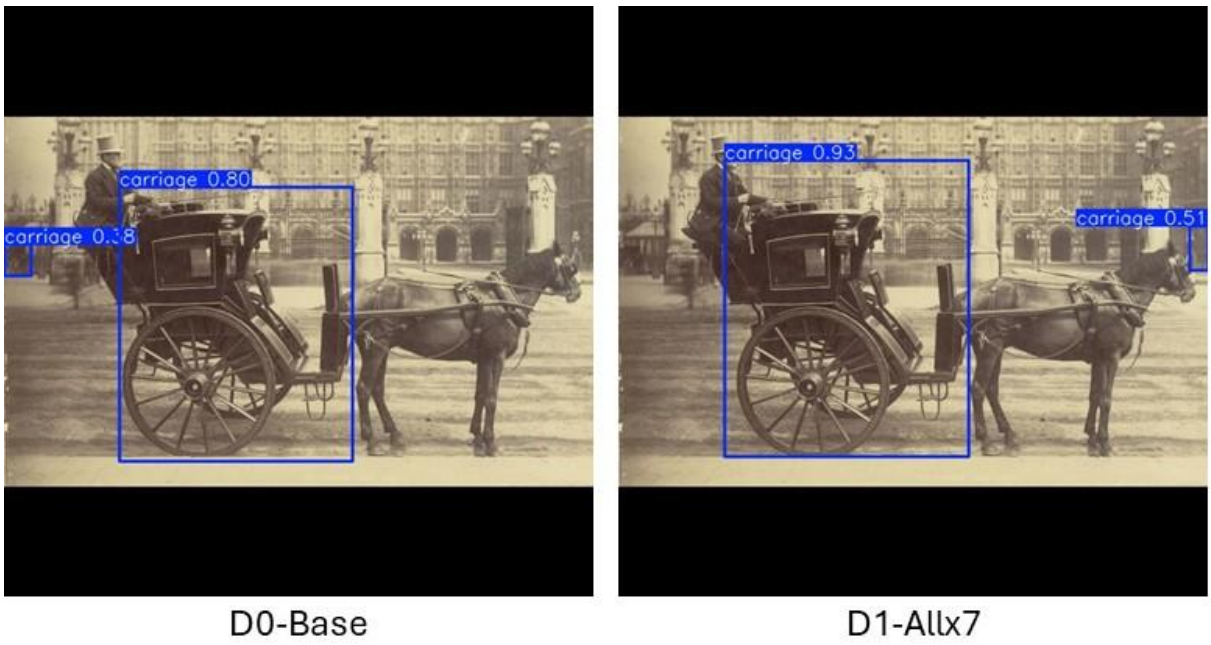


Figure 29: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: "The Hanson". Hästskjuts i London. *Photograph. Europeana (Swedish National Museum of Science and Technology). 1886. https://www.europeana.eu/nl/item/916118/S_TEK_photo_TEKA0156178.)*

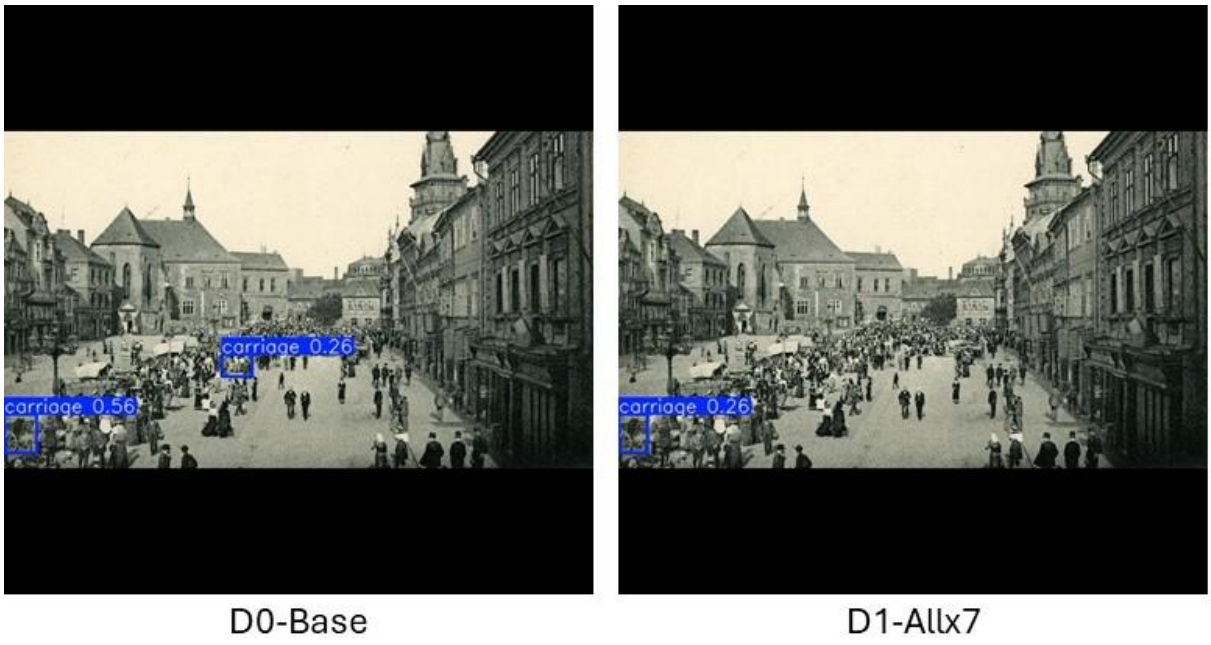


Figure 30: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Komotau. Marktplatz*. Photograph. Europeana (Deutsche Fotothek). 1912. https://www.europeana.eu/nl/item/440/item_W27ZVBJUP65WFSVBBKILFTB23FKO7ZHE.)*

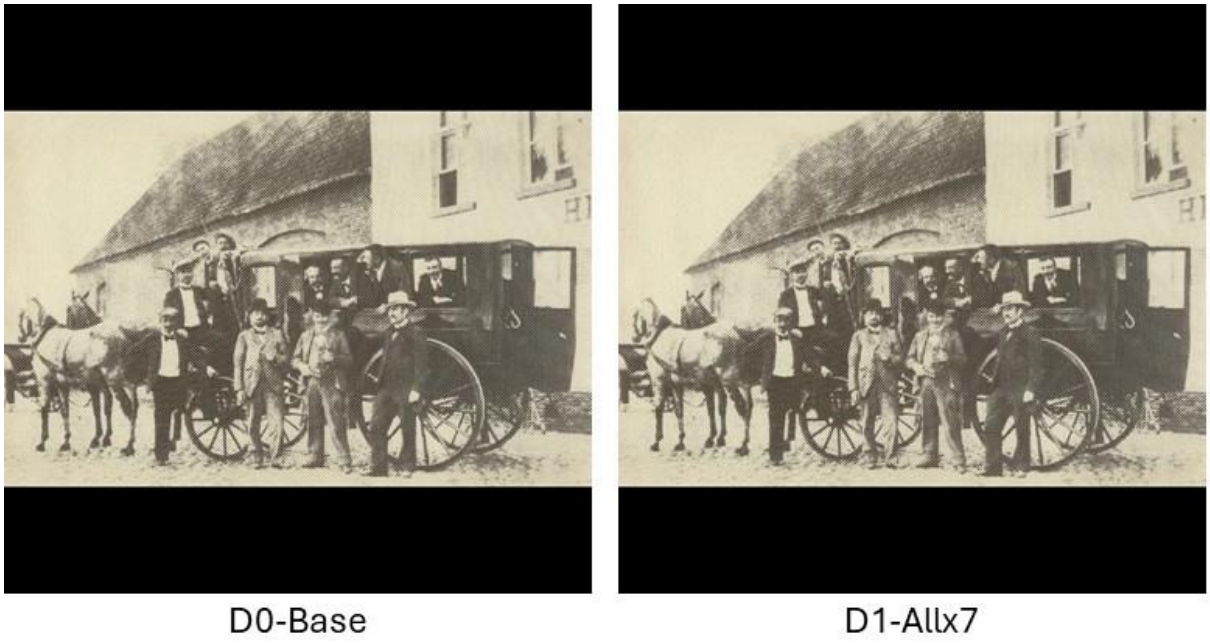


Figure 31: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Een koets vóór café Heidelberg te Zedelgem*. Photograph. ErfgoedBrugge. N.d. https://erfgoedbrugge.be/p/2300115_19_11987.)*

Figure 32: Comparison of results of YOLOv8 trained on different datasets containing carriages. (Source image: Leopoldplaats, het ruiterstandbeeld, Antwerpen. *Photograph. FelixArchief. 1910. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_4455&dtnr=1224_40&dtrecordid=29812&page=1&pageSize=10&type=copy.)*

6. Reimagining the past: CycleGAN as a tool for dataset expansion

As shown in the previous chapters, much of the current research in the heritage sector on object detection uses transfer learning, where pretrained object detection models are finetuned or retrained with a new dataset. While the new task of the model should be closely related to the task of the original model, the new dataset nevertheless requires the model to adapt. Even though, due to the similarity of the task, a pretrained model needs less data to be trained, it still needs a considerable amount of annotated data. The process of annotating can be very time-consuming and labour-intensive, depending on the complexity of the images and the constitution of the research group. Websites like Kaggle and Roboflow have online collections of free annotated datasets, which can be used for very domain specific tasks. Depending on the objective of the researchers, these datasets could be used to augment the training data and make the annotating work lighter. However, these datasets are mostly composed of contemporary images. This makes it less interesting when doing research based on historical pictures, because there is a stylistic gap that needs to be bridged.[128]

An interesting development in the world of computer vision and deep learning is style transferring. This is a method that enables the preservation of the core content of an input image while altering the visual style or appearance of the image to the style of another image.[129] In 2017 Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A. Efros published a paper where they presented CycleGAN, short for Cycle-Consistent Adversarial Networks.[130] This was a pivotal moment in the field of style transferring, because their proposed method made it possible to apply style transfer between unpaired images, something previously not possible.[131] Since the development of CycleGAN, it has been used frequently in research, with academics exploring the possibilities of this technique, also within the context of heritage. It is a new and relevant field, which becomes evident when looking at the body of literature used in this chapter, as many of the studies cited have been published within the last year.

Given the challenges posed by the scarcity of annotated historical datasets and the time-intensive process of manual annotation, CycleGAN offers an exciting opportunity to bridge the stylistic gap between modern and historical photographs. By making use of CycleGAN's ability to perform style transfer between unpaired datasets, contemporary images, which are often found in large, pre-annotated datasets, can be

---

[128] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 7.
[129] Zhaoxiang Tong, "Exploring the Impact of Hyperparameters on the Generation Quality of CycleGAN," *Transactions on Computer Science and Intelligent Systems Research* 5 (2024): 265, https://doi.org/10.62051/01m93a63.
[130] Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks."
[131] Bai, "Ancient Chinese Painting Style Transfer Based on CycleGAN," 130; Yen-Chia Chen et al., "Synthesis of Comic-Style Portraits Using Combination of CycleGAN and Pix2Pix," *Journal of Advanced Computational Intelligence and Intelligent Informatics* 28, no. 5 (2024): 1085, https://doi.org/10.20965/jaciii.2024.p1085.

transformed into historically styled images that are more aligned with the target dataset and augmenting this dataset for training an object recognition model. On the other hand, historical images could also be transformed into a modern style to facilitate their integration into existing models. This chapter seeks to explore the feasibility of this approach by addressing the following research question: “Can CycleGAN be used to alleviate annotation work for historical images by transforming existing modern annotated images to a historical style?”

6.1. Previous use of CycleGAN in a heritage context

Ever since its development, CycleGAN has been used in a myriad of research disciplines, one of them being related to heritage. In 2017, Vinson Luo, Michael Straka and Lucy Li used CycleGAN in a similar way as this chapter proposes, namely to transfer the style between historical and modern images.[132] The authors started with the observation that colourisation for historical photos can be challenging, because there is no ground truth for what the colours should be like.[133] In total, they used four datasets: one for historical images of people, one for their modern counterpart, one for historical photographs of landscapes and lastly one for modern photographs of landscapes.[134] First, the authors used the “standard” CycleGAN as developed by Zhu et al., but the results were not satisfactory. The model detected edges but was unable to colour objects reasonably and these colours were often inverted, so that lighter shades became black and darker shades became white. To tackle this problem, the authors slightly modified the architecture of their model to fit their needs. The authors discovered that, on their dataset, CycleGAN worked best on colouring objects with consistent colours. It thus performed best on the landscape dataset, where colours are indeed more consistent. When looking at the results of the people dataset, the researchers found two major flaws. On the one hand there were overly desaturated and sepia images and on the other hand there were overly bright and incoherent images. These are opposite of each other, which made it difficult to solve one without making the other worse.[135]

The research authored by Kim et al., already briefly discussed in the introduction, is also relevant in this context.[136] The researchers focused on the benefits of object detection in digital humanities, as they stated that right now, deep learning based object detection models are usually trained on large-scale datasets of real photos, which poses challenges for researchers in digital humanities.[137] The authors were

---

[132] Vinson Luo, Michael Straka, and Lucy Li, "Historical and Modern Image-to-Image Translation with Generative Adversarial Networks," (2017).
[133] Ibid., 1.
[134] Ibid., 3.
[135] Ibid., 7.
[136] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling."
[137] Ibid., 2-3.

interested in training an object detection model for identifying illustrated objects from nineteenth-century children's and youth literature. A problem that the authors noted, which is quite often the case for research within history and heritage studies, is the lack of annotated data to train the models.[138] To tackle this problem, the authors proposed an interesting method to collect annotated data.

First, the authors used a baseline model that was trained on a subset of the COCO 2014 training dataset and selected only the object classes that they needed. Thereafter, they used CycleGAN to transform COCO images into historical illustrations. By doing this, the authors could use these pre-annotated COCO images to train the model because the domain gap between the original COCO images and the historical illustrations was reduced. Finally, the researchers made use of Pseudo-Labels.[139] This is a semi-supervised learning technique that enables automatic annotation of objects.[140]

By using these techniques, the authors created four databases: a source dataset with the COCO images; a transformed dataset which contained the COCO images transformed by CycleGAN; and two "pseudo datasets" made up of the images with pseudo labels obtained by two different pretrained models, one being trained on the source dataset and the transformed dataset and the other on the source, transformed and first pseudo dataset.[141] These datasets were then used to train four models: one trained using the source dataset, one trained using the source dataset and the transformed dataset, one trained using the source dataset, the transformed dataset and the first pseudo dataset and finally one trained using the source dataset, the transformed dataset and the second pseudo dataset.[142] After testing, the authors concluded that the combination of domain transfer and pseudo labelling improved the performance of the object recognition model, which allowed the implementation of supervised learning without annotations on historical data.[143]

6.2. Results

For this research, a custom CycleGAN model was developed, by training a model on Google Colab with the L4 GPU for 200 epochs.[144] Two datasets were used for training, dataset A consisting of 780 modern photographs and dataset B consisting of 897 historical photographs. The training took around 8 hours.

---

[138] Ibid., 5-6.
[139] Ibid., 6-9.
[140] "Pseudo Labeling: Leveraging the Power of Self-Supervision in Machine Learning," Medium, updated 1 February 2024, 2024, accessed 27 December 2024, https://medium.com/@data-overload/pseudo-labeling-leveraging-the-power-of-self-supervision-in-machine-learning-d8192e918d65.
[141] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 8.
[142] Ibid., 9.
[143] Ibid., 8-9, 12-13.
[144] In the bibliography a Google Colab notebook can be found containing all the used code in this research.

In contrast to other machine learning models, the performance of CycleGAN is more subjective to measure.[145] During training, the three main loss functions, so the adversarial, identity and cycle consistency loss, were being monitored. The plots provide a little insight into the learning dynamics of the model, such as stabilisation and instabilities, as can be seen in Figure 33.[146] The most important curves here are the Cycle Consistency Losses, where Cycle A is showing the process from domain A (modern) to domain B (historical) and back to domain A and Cycle B the other way around, from domain B to domain A back to domain B. These plots illustrate how the model quickly adapted in the first 100 epochs (with the higher learning rate) and slowed down towards the end of the training (when the training rate went down each epoch). This loss reduction shows the model's advancement in comprehending and illustrating photographs from both domains.[147] When looking at the generator and discriminator losses, it becomes clear that they oscillate far more and don't really get to a plateau, contrary to the cycle consistency loss.

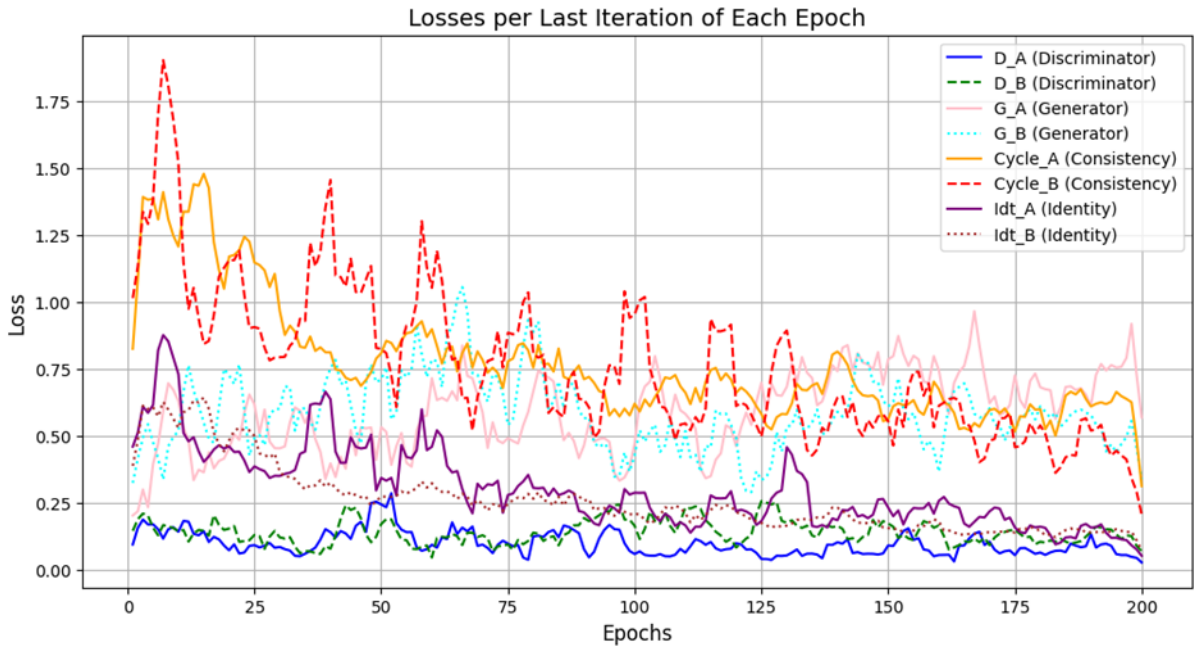


Figure 33: Training loss over 200 epochs, at the last iteration of each epoch.

However, as noted by the authors of CycleGAN, the loss curves alone do not reveal sufficient information about the quality of the results.[148] To truly measure how well the model worked, it is necessary to visually inspect the output of the model. After each epoch, the model saved eight images: the real image; the fake, generated image; the identity image and the reconstructed image. For evaluation of the trained

[145] Luo, Straka, and Li, "Historical and Modern Image-to-Image Translation with Generative Adversarial Networks," 5.
[146] The separate graphs of the different loss functions can be found in Annex 37.
[147] Hindarto, "Revolution in Image Data Collection: CycleGAN as a Dataset Generator," 451.
[148] Zhu, "pytorch-CycleGAN-and-pix2pix."

model in this chapter, the real and fake image were compared after a period of ten epochs to see how realistic the results are.

In Figure 34, the transformation of historical photographs into modern ones can be seen. Already in the early training phase, around epoch 30, consistent elements like the sky and ground began to show realistic colours, with skies turning blue and clouds showing white tones. Starting at epoch 50, more detailed elements, like buildings, started to be coloured, although there are some inconsistencies noticeable. The image at epoch 70 shows that the water and even the reflections in the water show logical and coherent colouring, which indicates the model's progress in understanding the more stable features in the photographs.

However, when looking at the later epochs, starting from epoch 100, more inconsistencies started to appear. While the more stable elements like the clouds and buildings are still realistic, there are some "artefacts" noticeable, for instance "blobs" of random colours occasionally emerged in the generated images. These irregularities disrupt the coherence of the transformation, which could be due to the model's limitations in handling less-predictable areas. A more significant issue arose with the depiction of people. For example, in a close-up photograph at epoch 160, a person became a dark shape, and at epoch 180, the people showed inconsistent and unrealistic colouring.

Epoch 1: Fake A

Epoch 1: Real B

Epoch 10: Fake A

Epoch 10: Real B

Epoch 20: Fake A

Epoch 20: Real B

Epoch 30: Fake A

Epoch 30: Real B

Epoch 40: Fake A

Epoch 40: Real B

Epoch 50: Fake A

Epoch 50: Real B

Epoch 60: Fake A

Epoch 60: Real B

Epoch 70: Fake A

Epoch 70: Real B

Epoch 80: Fake A

Epoch 80: Real B

Epoch 90: Fake A

Epoch 90: Real B

Epoch 100: Fake A

Epoch 100: Real B

Epoch 110: Fake A

Epoch 110: Real B

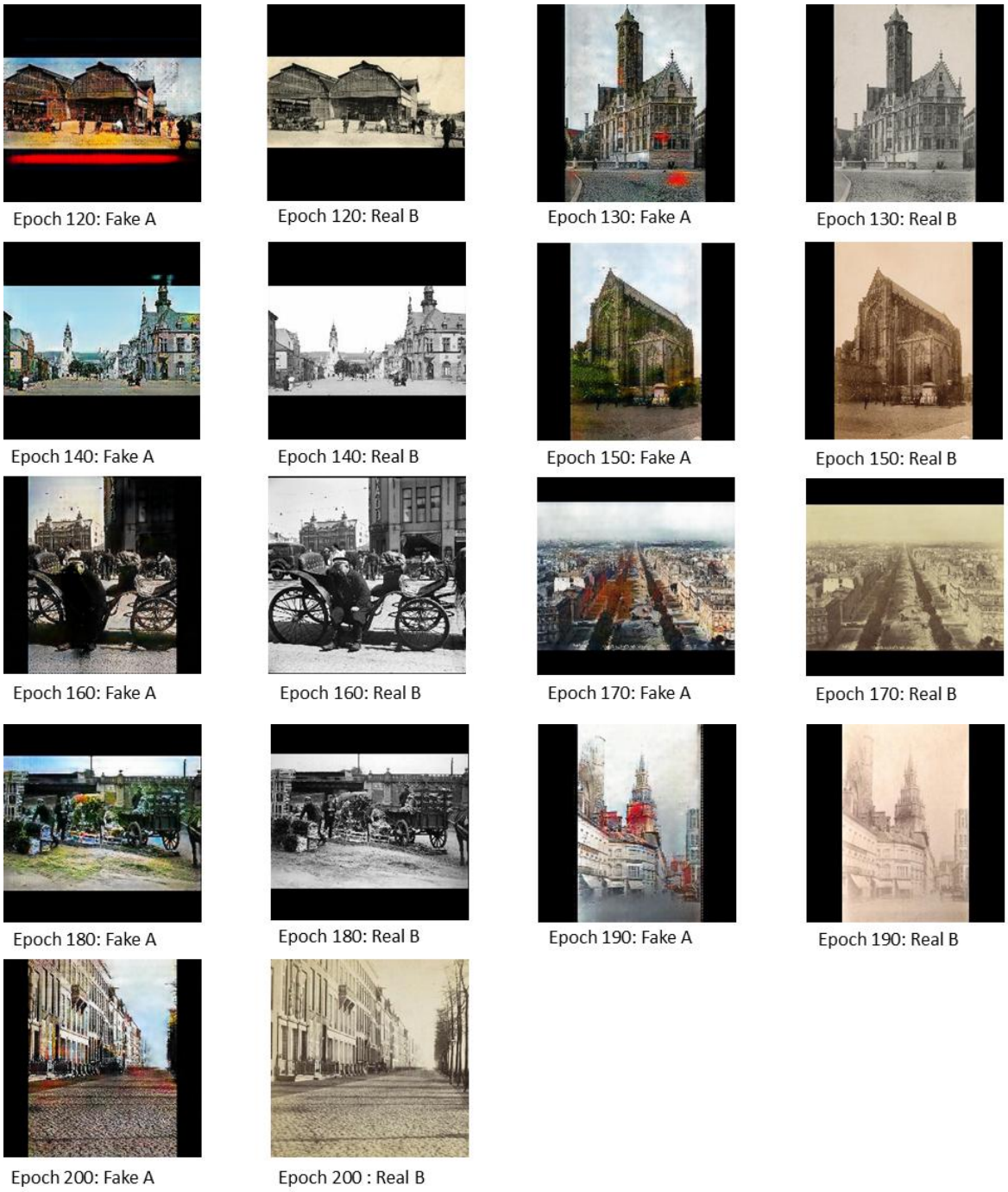


Figure 34: Progress per 10 epochs. Left the "fake" or created image, right the original image. (See bibliography for the sources of the images.)

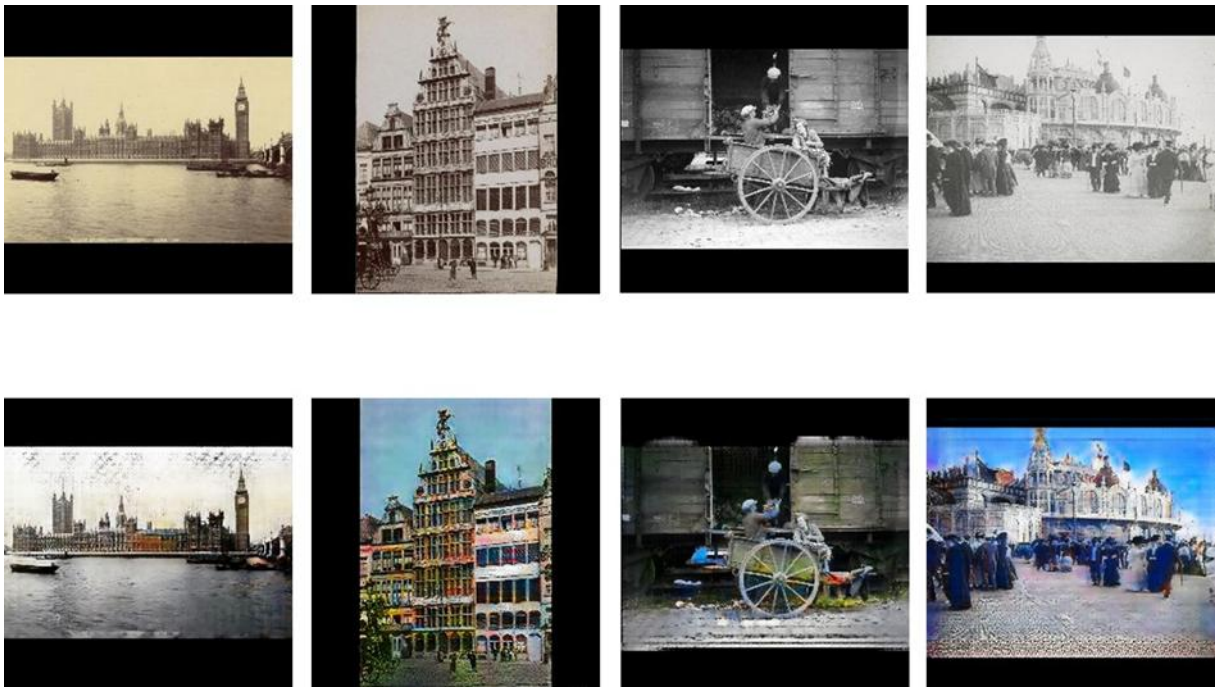


Figure 35: More detailed results after 200 epochs of training. At the top, the original, historical, images, at the bottom the "fake", modern, images. (See bibliography for the sources of the images.)

When looking at the transformation from modern to historical photographs in Figure 36, the model seems to perform much better, with the training also showing good results early on. The model quickly adapted to the monochromatic and sepia tones characteristic of historical photographs. Interesting to see here is that the model produced a mix of both black and white style images as well as more brown-toned images, which aligns well with the variety seen in the original historical dataset. This consistency in colour adaptation was maintained throughout the whole training process, which suggests a more reliable understanding of characteristics of historical photographs. Unlike the historical to modern transformation, the model handles the depiction of people more effectively. People are rendered with consistent and appropriate tones, and the overly bright or sepia artefacts that were noticed by Luo et al., are not seen in the own trained model.[149]

[149] Luo, Straka, and Li, "Historical and Modern Image-to-Image Translation with Generative Adversarial Networks," 7.

Epoch 1: Fake B

Epoch 1: Real A

Epoch 10: Fake B

Epoch 10: Real A

Epoch 20: Fake B

Epoch 20: Real A

Epoch 30: Fake B

Epoch 30: Real A

Epoch 40: Fake B

Epoch 40: Real A

Epoch 50: Fake B

Epoch 50: Real A

Epoch 60: Fake B

Epoch 60: Real A

Epoch 70: Fake B

Epoch 70: Real A

Epoch 80: Fake B

Epoch 80: Real A

Epoch 90: Fake B

Epoch 90: Real A

Epoch 100: Fake B

Epoch 100: Real A

Epoch 110: Fake B

Epoch 110: Real A

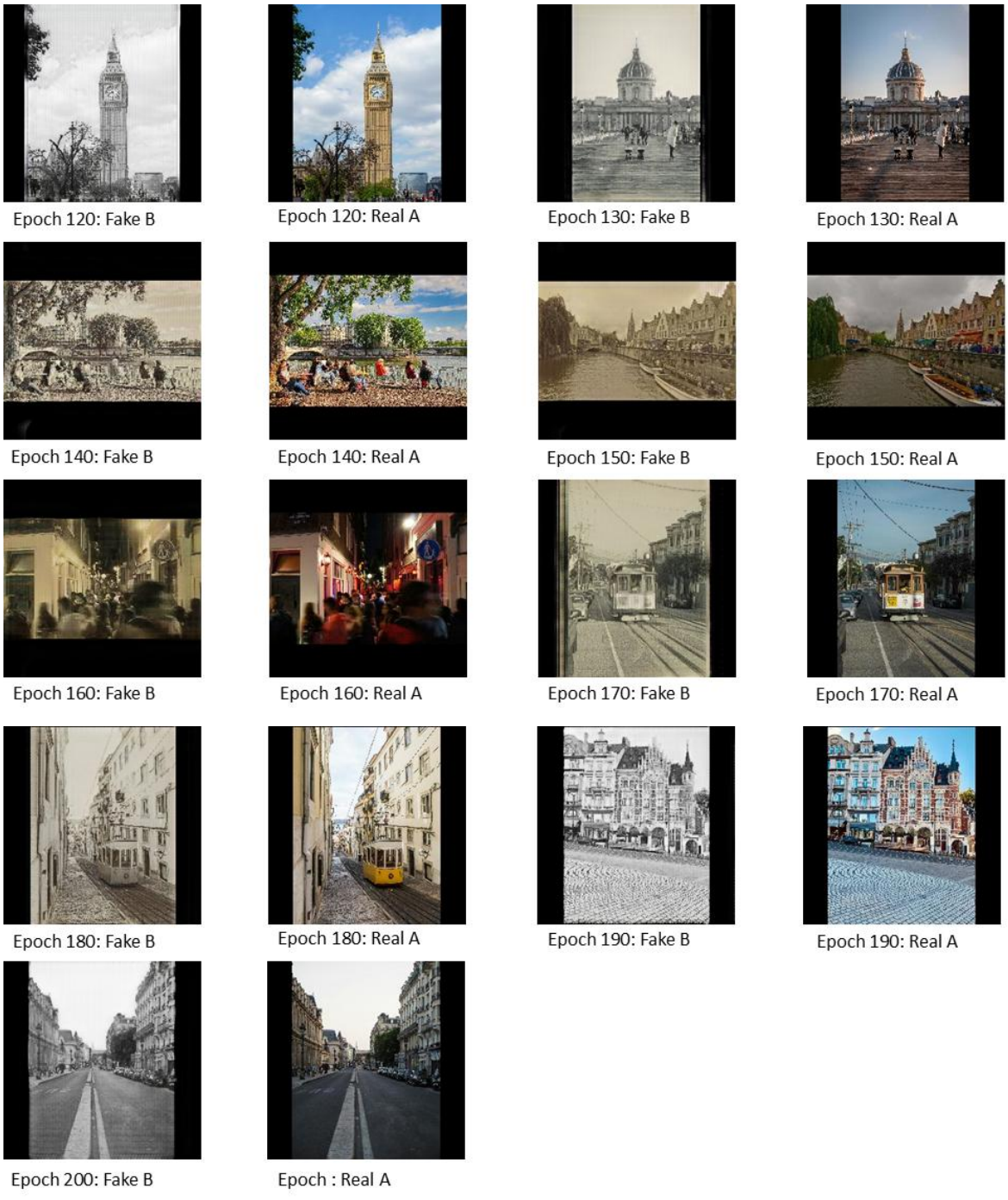


Figure 36: Progress per 10 epochs. Left the "fake" or created image, right the original image. (See bibliography for the sources of the images.)

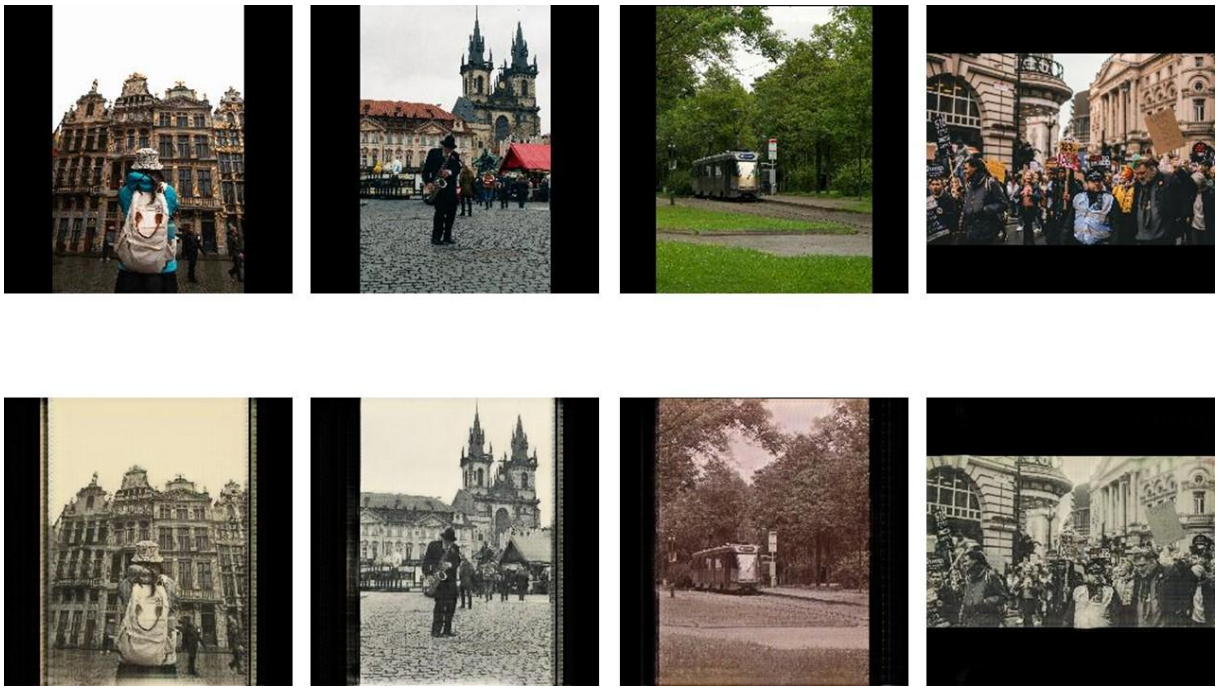


Figure 37: More detailed results after 200 epochs of training. At the top, the original, modern images, at the bottom the "fake", historical, images. (See bibliography for the sources of the images.)

The modern to historical transformation thus achieved better results overall, it produced consistent outputs that align closely to the real historical photographs. In contrast, the historical to modern transformation lagged a bit behind, particularly in handling dynamic elements like people. This discrepancy is understandable, since there is no “ground truth” present for historical photographs.[150] The model is thus forced to rely on less reliable predictions, which leads to greater inconsistencies. Looking back at the graph in Figure 33, this gap in performance between the two directions is also noticeable. The loss of GA, the one that is responsible for the transformation of photographs into the modern domain, is higher than the one of GB at the end of training. This indicates that GA does not produce images of the same quality as the ones produced by GB. Similarly, DB, which task it is to distinguish between real or fake images from domain B (historical), has a higher loss than DA, which has to decide whether an image from domain A (modern) is real or fake. This means that DB finds it more difficult to distinguish between real or fake images, pointing at the more realistic “fake historical” photographs produced by GA (modern to historical). Conversely, the lower loss of DA points toward the fact that this Discriminator finds it easier to establish whether an image is a real or fake modern one, which suggests that GB does not work as good.

[150] Ibid., 5.

7. Discussion

7.1. Discussion of the results

To answer the first sub-question of this research, “How do existing object detection models, such as YOLOv8, Faster R-CNN and RetinaNet, perform ‘out-of-the-box’ on historical and modern photographs?”, this study benchmarked these three object detection models on both modern and historical images. From this benchmark study, it could be concluded that YOLOv8l outperformed the other two models on both modern and historical datasets. It could also be seen that all models performed much better on modern images than on historical ones. This was expected, since historical images have different visual characteristics than their modern counterparts and since the models were trained on the Microsoft COCO dataset, consisting of modern images. The difficulties this domain gap pose, align with previous research of Ali and Zhang. They stated that: “Domain adaptability is another pressing challenge for YOLO. The model’s performance tends to degrade significantly in domains with irregular or noisy data, such as low-light environments, underwater imaging, or thermal imagery. These domains often present unique challenges that YOLO’s conventional architecture is not optimized to address, resulting in reduced detection accuracy and robustness.”[151]

To reduce this domain gap, YOLOv8l was subsequently trained on historical photographs containing trains and trams. This was done to answer the question “How does training an existing class in the model, such as trains/trams, using historical images affect the accuracy of the model?”. From the results, it became clear that the smallest dataset, containing 422 images of which 295 were used for training, did not affect the accuracy of the model in a meaningful way, only the recall improved. As expected, more data would be necessary to achieve an acceptable level of accuracy. Wei et al. pointed out that it often is not sufficient to simply only collect more new data, since newly gathered datasets inevitably introduce new variations such as lightning conditions, view angles,…[152] This is also true for historical photographs, which also vary due to factors like noise, lightning inconsistencies, and differences in colours such as sepia tones in one image and black-and-white in another. As a result, the domain gap remains a persistent challenge. As Ulmer et al. state, data augmentation has long been a crucial practice for training object detection models.[153] Shorten and Khoshgoftaar share this remark, stating that researchers often do not have access

[151] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection."
[152] Jian Wei, Qinzhao Wang, and Zixu Zhao, "YOLO-G: Improved YOLO for cross-domain object detection," *PLoS ONE* 18, no. 9 (2023): 1, https://doi.org/10.1371/journal.pone.0291241.
[153] Maximilian Ulmer et al., "How Important are Data Augmentations to Close the Domain Gap for Object Detection in Orbit?," *arXiv (Cornell University)* (2024): 4, https://doi.org/10.48550/arXiv.2410.15766.

to big data, and thus a useful solution to the problem of limited data is data augmentation.[154] Augmentation serves as a form of regularisation to prevent overfitting and it is also a means to increase the variance in the training data.[155] Carefully picked data augmentation strategies are thus important. For this research, several data augmentation techniques were chosen based on their relevance for historical photography. Compared to the benchmark results, small improvements in accuracy were made by the different augmentation strategies.

The results help to conclude that training an existing class with historical photographs does improve the accuracy of the model, especially the precision and recall, but that the performance gains were not immense, potentially due to the small dataset. Adding data augmentation to the dataset helps with improving the model further, but the improvement is not unlimited. There seems to be a tipping point beyond which additional data augmentation no longer leads to significant improvement. Shorten and Khoshgoftaar noted that combining different augmentations could lead to massively inflated dataset sizes, which is not always advantageous. In research with limited data, like this study, overinflating the dataset could even lead to further overfitting.[156] Furthermore, they also highlighted that no augmentation technique can correct an initial dataset that has poor diversity relative to the testing data, as the intrinsic bias in the original datasets limits generalisation.[157] Working with a smaller dataset means that it is more prone to bias due to limited variability and overrepresentation of specific viewpoints.

Since the objects covered by the 80 COCO classes are not always the objects a researcher wants to detect, custom classes often have to be added to an object detection model. This is the same for soundscape related research. There are many more objects that produce sound than those which appear in the COCO classes. Therefore, in the next chapter, a custom class was added to answer the research question: “How can an object detection model be extended to include a new class, such as carriages, and how well does the model perform on this new class?” The results showed that adding a new class to the object detection model is possible, but that even more data is needed for the model to perform well compared to a class that the model already knows. In line with the conclusions of Monna, it could be seen that the larger the dataset, the lower the gain in mAP became, because there already is more of the variability in types of objects in the larger dataset.[158]

---

[154] Connor Shorten and Taghi M Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data* 6, no. 1 (2019): 1, https://doi.org/10.1186/s40537-019-0197-0.
[155] Ulmer et al., "How Important are Data Augmentations to Close the Domain Gap for Object Detection in Orbit?."
[156] Shorten and Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," 17.
[157] Ibid., 42.
[158] Monna et al., "Deep learning to detect built cultural heritage from satellite imagery. - Spatial distribution and size of vernacular houses in Sumba, Indonesia," 178.

The dataset used for training a model on a custom class was also relatively small, which showed in the performance results. Although Wei et al. have stated that it does not suffice to only collect more data, the present dataset just seemed to be too small in the end and thus more new data would be necessary in the future.[159] As Mumuni et al. have pointed out: "The larger, more diverse and representative (with respect to the distribution of the target dataset) the training data, the more effective the deep learning model performs on unobserved data. On one hand, the diversity of training samples should be sufficient for the model to handle different deviations in image appearance and even noisy target instances. On the other hand, it is necessary that the quality does not degrade so much as to hamper performance on normal images."[160]

The results and visual representations from the two chapters in which a custom model was trained, showed that the different models did not always perform consistently. One of the major problems of complicated deep learning models like this, is the explainability of the results. It is very difficult to assess why the model did or did not detect a specific object, a phenomenon called the "black-box" problem.[161] The makers of YOLOv8, Ultralytics, mention that even AI engineers regularly have trouble with fully understanding the inner workings of their models.[162] This lack of transparency can make it harder to fine-tune models, since it is not sure which augmentation strategies have which effect on the final model.

The results discussed further highlighted the complex interplay between these data augmentation strategies and model performance. It could be seen that for the "train" class, a model like D6-Allx3 which combined all the augmentations, achieved the highest precision and mAP50 scores, while other models achieved the highest value for the other metrics. It seemed that no single model consistently outperformed all the others, which could both be seen in the metrical data as well as in the visual results. Ultimately, like the research of Kim also has shown, the "best model" depends on the purpose of the research.[163] If it is important to avoid false positives, a high precision should be prioritised, if the goal is to detect as many objects as possible, a high recall is more important and if overall accuracy across a range of thresholds is key, then a higher mAP50 or mAP50-95 should guide the choice.

---

[159] Wei, Wang, and Zhao, "YOLO-G: Improved YOLO for cross-domain object detection."
[160] Alhassan Mumuni and Fuseini Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array* 16 (2022): 1, https://doi.org/10.1016/j.array.2022.100258.
[161] Alain Andres et al., "On the black-box explainability of object detection models for safe and trustworthy industrial applications," *Results in Engineering* 24 (2024): 1, https://doi.org/10.1016/j.rineng.2024.103498.
[162] "All you need to know about explainable AI (XAI)," Ultralytics, 2024, accessed 17 April 2025, https://www.ultralytics.com/blog/all-you-need-to-know-about-explainable-ai.
[163] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 13.

Looking at the results of both chapters together, it can be seen that the same five challenges occur that Bengamra et al. saw with object detection in historical paintings.[164] The first challenge that they address, is the difficulty to detect objects depicted by different artists. While this is not exactly the same for historical photographs, the same reasoning still applies, namely that objects in different styles are more difficult to detect.[165] Looking at this research, it could be seen that horse-drawn trams were for example more difficult to detect than more contemporary looking trams. The same thing goes for the carriages. There were many different styles of carriages, some of which were less represented in the training data, causing the model to miss these on the unseen data. The visual diversity due to stylistic variances in the objects could help explain the performance of the models.

Next, they concluded that models had more difficulties with detecting objects in complex scenes.[166] As illustrated in this research, the own models also struggled with very busy or difficult scenes, for example when there were a lot of partially obstructed carriages on a marketplace. A third, related, challenge they noted, was the problem of occlusion in object detection.[167] When objects are partially obstructed by other objects, the model struggles to detect them. This is because these objects suddenly look different and do not correspond to the training data. In this research the same problem was noted, and in the benchmarking it became clear that YOLO seemed to suffer a bit more from this. In their study of the YOLO framework, Ali and Zhang also pointed toward the fact that the detection of small and occluded objects in complex environments is a major challenge for YOLO models, since they rely on anchor boxes and non-maximum suppression for object localisation, which can lead to false negatives and overlooked detections.[168]

The fourth problem they addressed, was the difficulty of detecting objects depicted from different viewpoints.[169] This is indeed a crucial problem, since this means that the training data has to be sufficiently diverse in terms of angles and perspectives in order for the model to generalise well on new data. Working with a smaller dataset in this thesis, achieving this diverse dataset was particularly challenging, which could have limited the model's performance. Finally, Bengamra et al. also stated that the scarcity of paintings annotated with bounding boxes around objects is a problem. This conclusion is shared by Marco Fiorucci et al., who state that the access to data and the quality of the data and metadata in the cultural

[164] Bengamra et al., "A comprehensive survey on object detection in Visual Art: taxonomy and challenge," 23-25.
[165] Ibid.
[166] Ibid.
[167] Ibid.
[168] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," 27.
[169] Bengamra et al., "A comprehensive survey on object detection in Visual Art: taxonomy and challenge."

heritage field is far from perfect, although institutions seem to be working hard to address this problem.[170] In this thesis, the same observation was made. Platforms like Europeana greatly help with the collection of data, but the absence of an annotated dataset containing historical pictures is a limiting factor for this type of research, since this has to be done manually which takes a lot of time.

Since CycleGAN is being used more and more as an efficient data augmentation technique, the last sub-question this research set out to explore was: "Can CycleGAN be used to alleviate annotation work for historical images by transforming existing modern annotated images to a historical style?"[171] This was done by training a custom CycleGAN model with a dataset containing modern photographs and one containing comparable historical photographs. The results were mixed. The custom model performed well when translating modern images into a historical style, but it was less consistent in the opposite direction, so from historical to modern. Contrary to previous research by Luo et al., where they found that the transfer from modern to historical resulted in overly desaturated, sepia-toned images or overly bright and incoherent results, the own model did not seem to suffer from these problems.[172] However, the style transfer from historical to modern images showed inconsistent outcomes, with drawbacks like unrealistic colours and the appearance of artefacts. This aligns with the findings by Pai et al., who point out that the presence of artefacts is a known limitation of GANs.[173] Park also noted that while CycleGAN can effectively learn to transfer style between unpaired images, it often struggles to accurately preserve content information such as the colour of individual objects, which lead to artefact formation and unrealistic colour shifts.[174]

7.2. Discussion of the implications

Combining all these results allows for answering the central research question of this study: "How can machine learning be used to detect sound sources in historical photographs?" This pilot study set out to explore this research question in four phases: first by benchmarking object detection models and selecting the best performing one; then training the best one on a known class, "train"; subsequently extending the model to include a new class, "carriage", a step that will undoubtedly be very necessary in future research on historical soundscapes; and finally by exploring machine learning in the form of CycleGAN to

---

[170] Marco Fiorucci et al., "Machine Learning for Cultural Heritage: A Survey," *Pattern Recognition Letters* 133 (2020): 103, 06-07, https://doi.org/10.1016/j.patrec.2020.02.017.
[171] Hindarto, "Revolution in Image Data Collection: CycleGAN as a Dataset Generator," 444-45.
[172] Luo, Straka, and Li, "Historical and Modern Image-to-Image Translation with Generative Adversarial Networks," 7.
[173] Suraj Pai et al., "Frequency-Domain-Based Structure Losses for CycleGAN-Based Cone-Beam Computed Tomography Translation," *Sensors* 23, no. 3 (2023), https://doi.org/10.3390/s23031089.
[174] Jaihyun Park, David K. Han, and Hanseok Ko, "Adaptive Weighted Multi-Discriminator CycleGAN for Underwater Image Enhancement," *Journal of Marine Science and Engineering* 7, no. 7 (2019), https://doi.org/10.3390/jmse7070200.

artificially increase the training data. The results from the different chapters indicate that training a custom model to detect sound related objects in historical photographs is a promising direction, but that a substantial amount of diverse and annotated data is required to achieve good performance. In the own study, the models were trained with relatively small datasets, which limited strong conclusions. Nevertheless, the results did suggest the promise of machine learning as a tool for data collection in soundscape research.

In an ideal world, these detections could be linked to the metadata of the photograph, which would help researchers to detect when and where a certain sound could be heard. However, this method is only one of the multiple ways to collect data for historical soundscape reconstruction. As stated in the introduction of this research, there already are different studies conducted which use machine learning on textual sources. A combination of these two methods could offer a more comprehensive and reliable approach to gather sources for soundscape reconstruction. Where object detection can identify physical sources of sounds, this can be complemented by textual sources that mention how these sounds were experienced or perceived. This is however only the very first step in the process of reconstructing soundscapes. As Pardoen illustrated, building a reliable soundscape requires much more contextual and technical work after the sources are collected.[175]

Another outcome of this research is the relative ease of training a custom object detection model. As Monna noted: “A novice might be afraid of the technicity required and the potential difficulties related to practical implementations. It would however be an error because all tools, freely available, come with detailed documentation.”[176] Thanks to the rapid evolving field, the training and application of custom machine learning models is becoming easier. Platforms like Roboflow help users through the whole process of training a model by providing them with an intuitive interface and reducing the need for extensive coding knowledge. During this project, the platform got a few more updates, becoming even more user-friendly. Lee also stated that in heritage fields the sharing of code is very important, since often projects are undertaken without professional software engineers.[177] It is thanks to this shared knowledge, like the prebuilt Colab notebooks that were used in this project, that these kinds of projects are becoming possible for people not trained as computer scientists.

---

[175] Pardoen, "Projet Bretez: une pincée de son dans l’Histoire.," 13-15.
[176] Monna et al., "Deep learning to detect built cultural heritage from satellite imagery. - Spatial distribution and size of vernacular houses in Sumba, Indonesia," 181.
[177] Benjamin Charles Germain Lee, "The “Collections as ML Data” checklist for machine learning and cultural heritage," *Journal of the Association for Information Science and Technology* 76, no. 2 (2023): 389-90, https://doi.org/10.1002/asi.24765. In the bibliography a Google Colab notebook can be found containing all the used code in this research.

Despite heritage professionals being trained as jack of all trades, fully understanding the model is often difficult and, as will be discussed in the next section, there are also a lot of technical possibilities with the training of a custom model, like hyperparameter tuning and changing parts of the model architecture. These often require very specific knowledge, and that is why an interdisciplinary approach should be the golden standard for this kind of research. Like many of the studies in this research show, it is very often a combination of research from the humanities field and computer scientists that work together.

It is important to also take into account the environmental impact of using deep learning methods in research. As Ali and Zhang stated: "[…] while the technology itself is seen as cutting-edge, the infrastructure required to support its deployment is resource-intensive, contributing to the carbon footprint of AI technologies. […] Datacenters, GPUs, and cloud computing infrastructures essential for training large-scale models consume vast amounts of energy. Therefore, it is essential that we focus on developing greener AI technologies and optimizing YOLO variants for energy efficiency.". [178] The question whether such approaches based on deep learning are always necessary is a justified one therefore. A balance needs to be found between the environmental impact of the project and the time savings achieved by employing a model on the research. This is a reason for heritage organisations and scholars to join forces and develop one well-performing model trained on historical photographs. Such a model could detect a wide range of objects and be repurposed for various applications, from soundscape reconstructions to archival annotations. A project like that prevents that different researchers and institutions are going to develop own models, leading to a more sustainable future and possibly a better performing model.

The implications of the results of the CycleGAN model are twofold. CycleGAN can be applied to transform modern photographs into historical looking ones. In that way, the dataset gets augmented and there is less annotating work, because pre-annotated modern images can be used for transformation. Since CycleGAN only changes the style of the images and keeps the core content, the annotations in the form of bounding boxes still stay the same. But CycleGAN can also be used the other way around, from historical photographs to modern ones. By performing this stylistic transformation, the stylistic gap existing in pretrained models could be bridged. The pretrained models are indeed trained on modern datasets, like the COCO dataset. This means that the model might perform worse on historical photographs. But if this historical dataset is transformed with CycleGAN to resemble modern photographs, it would theoretically not be necessary to retrain the model. Doing this would provide useful insights in how adaptable models are across time periods. Still, as shown before, this last option is more difficult. As Doan et al. stated:

[178] Ali and Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," 31.

“Furthermore, image-to-image translation is a widely used approach in UDA for object detection. It uses the GAN principle to transform the appearance of images from the target domain to resemble the source domain. However, training a GAN-based model is a complex task due to its minimax optimisation problem. This can lead to various challenges such as non-convergence, mode collapse, and diminished gradient.[179]

7.3. Discussion of the limitations and future research

Being a pilot study, this research faced many limitations and has different opportunities for further research. As pointed out throughout this research, the experiments and training were conducted with relatively small datasets. When combining all the different, unaugmented, datasets, nearly 1600 historical photographs were manually annotated, something that took a considerable amount of time. By using different data augmentation techniques, these datasets were expanded, but, as stated in the previous section, data augmentation can only go so far in improving the performance of an object detection model. Compared to other research, like that of Kim et al. where the largest dataset contained over 100000 images, the own datasets are relatively small.[180]

Further research could thus experiment with larger datasets. Here, the use of crowdsourcing methods to annotate more photographs could be a useful tool. For the scene detection model trained on the *De Boer* press photo collection, the Noord-Hollands Archief used the help of 652 participants to match tens of thousands of photographs with a description of the event depicted.[181] A similar project could be created where people get a set of historical photographs and have to draw bounding boxes around the desired objects. In that way, a large-scale dataset with historical photographs that could also be used for other projects than soundscape research that also use historical photographs, would be interesting. Since historical photographs come with different issues, like large viewpoint differences, digitisation artefacts and unknown camera parameters, Ferdinand Maiwald urged already in 2019 for the establishment of a benchmark dataset consisting of different historical photographs.[182] Although he wanted to use the dataset for evaluating feature matching methods, it could be interesting to expand such initiatives to use it for more diverse projects. In the field of art sciences, similar initiatives where annotated datasets for machine learning are created already exist. Sticking with the theme of this research for example, the Medieval Musicological Studies Dataset is an annotated dataset containing medieval artworks used for

[179] Anh-Dzung Doan et al., "Assessing domain gap for continual domain adaptation in object detection," *Computer Vision and Image Understanding* 238 (2024): 2, https://doi.org/10.1016/j.cviu.2023.103885.
[180] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 9.
[181] "Fotografisch Geheugen: de reportages van Fotopersbureau De Boer," VeleHanden, accessed 22 April 2025, https://velehanden.nl/projecten/bekijk/details/project/ranh_tagselection_deboer.
[182] Ferdinand Maiwald, "Generation of a benchmark dataset using historical photographs for automated evaluation of different feature matching methods," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, no. 2 (2019), https://doi.org/10.5194/isprs-archives-XLII-2-W13-87-2019.

object detection with connection to musical instruments and singing poses.[183] At the same time, it is also important that the providing heritage institutions add as much metadata as possible to the photographs, since for projects like this, it is important to know when and where the photographs are taken.

Next to the data augmentation techniques discussed in this research, it could also be interesting to explore how synthetic images created by generative image models, like OpenAI's DALL·E, could enhance a dataset. There already have been a few pioneering studies in other domains making use of these models, which offered promising results.[184] Just for exploring the potential of DALL·E 3, two synthetic images were generated here. First, a text-to-image method was tested by providing the model with the following prompt: "A realistic historical photograph of a horse-drawn tram moving through a late 19th-century European city street. The scene is lively, with people wearing late 19th-century European clothing. The image is sepia-toned, with visible film grain, slight blur, and subtle aging effects like scratches and faded edges. The style imitates authentic archival photography from the 1800s." Next, this prompt was reused, combined with a reference image to guide the model better, which is called image-to-image generation. This resulted in Figure 38 and Figure 39. The resulting synthetic images are interesting, but they also contain some errors. In both figures there are for example three train tracks, where there should be two or four and in Figure 39 the horse-drawn tram also is "electrified" at the top. Future research could then explore how this technique compares to traditional data augmentation techniques.

[183] Bekkouch Imad Eddine Ibrahim et al., "Few-Shot Object Detection: Application to Medieval Musicological Studies," *Journal of Imaging* 8, no. 18 (2022), https://doi.org/10.3390/jimaging8020018; Bengamra et al., "A comprehensive survey on object detection in Visual Art: taxonomy and challenge," 9.

[184] Ranjan Sapkota and Manoj Karkee, "Generative AI in Agriculture: Creating Image Datasets Using DALL.E's Advanced Large Language Model Capabilities," *arXiv (Cornell University)* (2023), https://doi.org/10.48550/arXiv.2307.08789; Yuwei Yin et al., "TTIDA: Controllable Generative Data Augmentation via Text-to-Text and Text-to-Image Models," ibid., https://doi.org/10.48550/arXiv.2304.08821; "Launch: Synthetic Image Generation with DALL-E and GPT-4 Vision," Roboflow, 2023, accessed 3 May 2025, https://blog.roboflow.com/synthetic-data-dall-e-roboflow/; Jeongmin Shin and Hyeryung Jang, "Data Augmentation Techniques Using Text-to-Image Diffusion Models for Enhanced Data Diversity" (2024 15th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, IEEE, 2024).

Figure 38: Synthetic image created with DALL·E by using a custom prompt.

Figure 39: Synthetic image (left) created with DALL·E by using a custom prompt and providing it with a reference image (right). (Source image: Stadhuis Antwerpen*. Photograph. FelixArchief. C. 1870-1882. https://felixarchief.antwerpen.be/detailpagina?invnr=934_21909&dtnr=1224_40&dtrecordid=110084&page=1&pageSize=10&type=copy.)*

Another limitation of this exploratory research is the fact that it focused on training and testing one object class at a time. This does not fully reflect the ultimate goal of sound source detection in historical photographs, as this will need an object detection model that detects multiple relevant classes simultaneously. This means that all relevant objects on photographs need to be annotated for training a model, which was not feasible for this project. Important to note here is also the problem of imbalanced datasets, where some classes have more instances than other. As seen in the benchmark dataset, class imbalance is also a very relevant problem for historical photographs, since some classes, like “people”, just naturally appear more than others. During model training, the majority classes often gets a strong bias, which leads to a model that could achieve a high overall accuracy but fails to effectively detect items

from the minority classes.[185] The impact of this problem is more pronounced in single-stage detectors, like YOLO.[186] Since only one class was tested in this research, it remains uncertain how well a model like YOLOv8l performs on more classes when retrained. Future research could add more classes to the model, and test techniques to counter class imbalances, like oversampling or augmentation of the minority classes by using CycleGAN.[187]

Additionally, while small improvements were noted through training on historical images, and a new class was added to the model, the overall performance of the object detection model is not yet satisfactory. This can be attributed to the limited data that was used to train the model, but also due to the fact that there were no experiments with hyperparameter tuning. During training, hyperparameters, like learning rate, batch size, number of epochs or architecture specific details, could be tuned to tweak the performance of the model.[188] In principle, hyperparameter tuning is not a one-time set-up, but it is an iterative process where the best settings for the hyperparameters are searched.[189] The problem here is that this can be a very resource intensive process, since models need to be retrained each time.[190] Faced with limited computational resources due to the costly nature of Google Colab Pro, there was little opportunity for this research to experiment with different hyperparameters, which could have improved results. The use of more lightweight models, such as YOLOv8n, might reduce resource requirements in future work, but this comes at the cost of decreased accuracy. Furthermore, the standard settings for resolution (640x640) were used. But this could potentially be made higher, as Shorten and Khoshgoftaar not: “Many current models downsample images from their original resolution to make the classification problem computationally more feasible. However, sometimes this downsampling causes information loss within the image, making image recognition more difficult.”.[191]

The same things apply for CycleGAN. While the use of CycleGAN has much potential and the transformation from modern to historical style images was already good, it was illustrated that the model still struggled with some transformations from historical to modern style images. Zhaoxiang Tong explored the different possibilities with hyperparameter tuning for CycleGAN, and concluded that, based on the aim of the research project, setting the hyperparameters to specific values had a profound

---

[185] Kim, Im, and Mandl, "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling," 9.
[186] Nieves Crasto, "Class Imbalance in Object Detection: An Experimental Diagnosis and Study of Mitigation Strategies," *arXiv (Cornell University)* (2024): 1, https://doi.org/10.48550/arXiv.2403.07113.
[187] Shorten and Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," 18-19.
[188] "Ultralytics YOLO Hyperparameter Tuning Guide," Ultralytics, accessed 24 April 2025, https://docs.ultralytics.com/guides/hyperparameter-tuning/.
[189] Ibid.
[190] Ulmer et al., "How Important are Data Augmentations to Close the Domain Gap for Object Detection in Orbit?," 5.
[191] Shorten and Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," 38.

impact.[192] Future research could thus explore with a larger dataset and different hyperparameters. Moreover, in this research, the input resolution of the images was set to 256x256 pixels. This was done because CycleGAN is already very resource intensive, and in that way training could proceed faster. The problem however is that resizing means that some details get lost, which could cause problems when performing object recognition. Of course the resolution could be made higher, it is fairly easy to change the code so that CycleGAN works with higher quality images, but this has then again consequences for the time and resources it takes for training the model. Shorten and Khoshgoftaar also stated that a higher resolution is more difficult: "Resolution is also a very important topic with GANs. Producing high resolution outputs from GANs is very difficult due to issues with training stability and mode collapse."[193]

Finally, in the benchmarking test, only three object detection models were evaluated, namely YOLOv8, Faster R-CNN, and RetinaNet. This selection was based on the user-friendliness of the models and their popularity, and it offered a first exploration of the difference between performance of one-stage and two-stage detectors on historical photography. These models are however only the tip of the iceberg. There are many different architectures available, and new ones get released very often. A larger comparative study could perhaps provide different insights and discover models more suited for the style of historical photographs.

[192] Tong, "Exploring the Impact of Hyperparameters on the Generation Quality of CycleGAN."
[193] Shorten and Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," 39.

8. Conclusion

With the growing interest in historical soundscapes on the one hand, and the rapid evolvement of machine learning on the other, new and innovative methods for data collection in heritage studies are emerging. One little studied subject, however, is the performance of object detection on historical photographs. This thesis set out to explore the feasibility of training an object recognition model to automatically detect sound sources on historical photographs, in order to use it as a way to collect data for historical soundscape reconstruction. The main research question of this thesis was: "How can machine learning be used to detect sound sources in historical photographs?". This question broke down into several sub-questions: 1) How do existing object detection models, such as YOLOv8, Faster R-CNN and RetinaNet, perform 'out of the box' on historical and modern photographs? 2) How does training an existing class in the model, such as trains/trams, using historical images affect the accuracy of the model? 3) How can an object detection model be extended to include a new class, such as carriages, and how well does the model perform on this new class? 4) Can CycleGAN be used to alleviate annotation work for historical images by transforming existing modern annotated images to a historical style?

The results of this thesis demonstrated both the potential as well as the drawbacks of using object detection on historical photographs. Testing three different object detection models 'out of the box' on modern and historical photographs showed that all of these models performed noticeably worse on historical material due to the domain gap, consisting of visual and stylistic differences. This confirmed that 'out of the box' models pretrained on datasets containing modern photographs, like the COCO dataset, are not immediately suited for historical sources and that fine-tuning on a custom dataset was thus necessary. YOLOv8l seemed to be the model least influenced by the stylistic differences and was therefore fine-tuned on a training dataset consisting of historical photographs to reduce this domain gap.

First, this was done by training a model on one of the 80 COCO classes, "train". While the smallest dataset saw an improvement in recall, its precision went down. The dataset was therefore augmented using different techniques, to see how these techniques influenced the performance. Compared with the smallest dataset, nearly all of the different data augmentation techniques had a positive effect, with a combination of all the different techniques proving to be the more balanced option. Adding even more data augmentation was not necessarily beneficial, the law of diminishing returns applied here. In the end, only modest improvements were made compared to the benchmarking test. While the precision and recall improved with about 10%, the mAP only improved marginally. To see how easy the model performed on unfamiliar territory, a custom class, "carriage", was added to the model. As expected, the model performed worse on this new class, even with double the amount of training instances compared to a "known" class. Several interesting results emerged nevertheless, from the difficulty posed by

variability in visual representations of carriages, over the scarcity of annotated training images, to the challenge of detecting objects in complex or occluded scenes.

Finally, this thesis also explored the use of CycleGAN to make modern photographs look like historical ones and vice versa. This could theoretically reduce manual annotation needs, since on the one hand pre-annotated modern photographs could be translated into historical ones to use them as training data and on the other hand historical photographs could be made modern and could then directly be used in a pretrained object detection model. The results showed that the direction of modern to historical is promising, since the generated images were visually convincing and avoided many of the issues described in pervious studies. However, translating historical images into modern ones proved more problematic, since the model often introduced random artefacts and inconsistencies, which corresponded to previous research showing that training a CycleGAN model is complex and not always stable.

The pioneering nature of this thesis also caused several limitations that should be acknowledged. Most important here are the relatively small training datasets used to fine-tune the object detection model and the limited computational resources. As shown in the results, the training datasets were on the smaller side which limited strong conclusions or high-grade performance of the models. Larger datasets, containing new, “fresh”, data and not only augmented data, will improve the models significantly. The limited computational resources made it not possible to explore hyperparameter tuning and explore with higher quality images, both in the object detection models as well as in the custom CycleGAN model. However, with these limitations also come many exciting possibilities for further research. Next to the experimentation with different hyperparameters and the possibility of using crowdsource methods to develop large annotated datasets, the use of generative models like DALL·E and the expansion of model benchmarking to include newer architectures and more classes, seem interesting avenues.

While the relatively small datasets and the limited computational resources limited the generalisability of the results, this thesis nevertheless showed that object detection on historical photographs for historical soundscape reconstruction, and perhaps more broadly for other heritage and historical related research, could be a valuable and promising tool. If researchers from various disciplines, like heritage studies, history and computer science, join forces, it becomes possible to develop a sturdy model tailored to the needs of cultural and historical research. This will allow for automating parts of the data collection process, which means saving time and being able to effectively analyse larger collections of archival sources. While such a model will not include every relevant class required by different projects, the creation of a shared base model trained on a large and varied historical dataset could serve as a valuable resource, since it can then be fine-tuned for specific research. This thesis contributed to filling a part of the blind spot in the current research on this topic. It is hoped that this study serves as an initial impetus

for further research that builds on these foundations. As tools, datasets, and collaboration continue to evolve, so too will the possibilities for machine learning in heritage studies.

Bibliography

Primary sources

Dataset

- Historical
    - "Erfgoed Brugge." Erfgoed Brugge, accessed 10 October 2024, https://erfgoedbrugge.be/.
    - "Erfgoedbank Brussel." Erfgoedbank Brussel, accessed 10 October 2024, https://erfgoedbankbrussel.be/.
    - "Europeana." Europeana, accessed 10 October 2024, https://www.europeana.eu/nl.
    - "Felixarchief." Stad Antwerpen, accessed 10 October 2024, https://felixarchief.antwerpen.be/.
    - "Library of Congress." Library of Congress, accessed 10 October 2024, https://www.loc.gov/.
- Modern
    - "Unsplash." Unsplash, accessed 10 October 2024, https://unsplash.com/.

Cover Image

- *Grote Markt met vrijheidsboom voor het stadhuis*. Photograph. FelixArchief. C. 1920-1960. https://felixarchief.antwerpen.be/detailpagina?invnr=752_175&dtnr=1224_62&dtrecordid=2695&page=1&pageSize=10&type=copy.

Chapter 2

- Figure 1:
    - Real Image in domain A: Peeters, Florian. *A city street with cars parked on both sides*. Photograph. Unsplash. June 19, 2023. https://unsplash.com/photos/a-city-street-with-cars-parked-on-both-sides-gBWhIm1JPB4.

    - Real Image in domain B: *Boompjes Rotterdam*. Photograph. Europeana (Rijksmuseum). 1860-1880. https://www.europeana.eu/nl/item/90402/RP_F_F12132.

- Figure 2: *Budapest. Klinika Üllöi út (1877)*. Photograph. Europeana (Deutsche Fotothek). 1877. https://www.europeana.eu/nl/item/440/item_PWWSSBUGOLDEHDTPOAP52YOA76NW4DH2.

- Figure 3: *De Keyserlei: het Centraal Station, Antwerpen 1910*. Photograph. FelixArchief. 1910. https://felixarchief.antwerpen.be/detailpagina?invnr=PB_2005&dtnr=1224_40&dtrecordid=22623&page=1&pageS.

Chapter 3

- Figure 5: *Hästvagn*. Photograph. Europeana (Malmö Museum). N.d. https://www.europeana.eu/nl/item/91672/MM_foto_612340.

- Figure 6: *Weltausstellung Paris 1867*. Photograph. Europeana (Albertina). 1867. https://www.europeana.eu/nl/item/15508/FotoGLV2000_22294.

- Figure 7: *Stadsgata*. Photograph. Europeana (Blekinge Museum). 1900-1915. https://www.europeana.eu/nl/item/916119/blm_item_68208.

- Figure 8: *Ata, hästvagn*. Photograph. Europeana (Malmö Museum). 1910. https://www.europeana.eu/nl/item/91672/MM_foto_600289.

- Figure 9: *Wien Hofoper*. Photograph. Europeana (Albertina). 1908-1909. https://www.europeana.eu/nl/item/15508/FotoGLV2000_8382.

- Figure 10: *Wien, Szene mit Straßenbahn*. Photograph. Europeana (Albertina). 1906-1907. https://www.europeana.eu/nl/item/15508/Foto2005_177_32.

- Figure 11: *Spårvagn, hästspårvagn*. Photograph. Europeana (Malmö Museum). 1907. https://www.europeana.eu/nl/item/91672/MM_foto_611540.

- Figure 12: *Gezicht op de Jutfaseweg te Utrecht met de paardentram Utrecht-Jutfaas/Vreeswijk*. Photograph. Europeana (Het Utrechts Archief). 1895. https://www.europeana.eu/nl/item/257/https___hetutrechtsarchief_nl_beeld_CBC12DD377C95F9E94E1FC6A5F2470FA.

Chapter 4

- Figure 14: *Suikerrui, vanaf het Zuiderterras, Antwerpen*. Photograph. FelixArchief. 1909. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_7087&dtnr=1224_40&dtrecordid=32372&page=1&pageSize=10&type=copy.

- Figure 15: *Gezicht op het S.S.-station Den Haag S.S. te Den Haag*. Photograph. Europeana (Het Utrechts Archief). 1899. https://www.europeana.eu/nl/item/257/https___hetutrechtsarchief_nl_beeld_025EB5A6A32053F7A8961CD4B2798CA9.

- Figure 16: *Sint-Agatha-Berchem: Grand'Place en Gentsesteenweg met tram*. Photographs. Erfgoedbank Brussel. N.d. https://erfgoedbankbrussel.be/mediabank/detail/f8619c70-3828-80b1-d9bb-17fc6a401c34/media/4876b394-3fe3-fb1c-c95b-618d29dd33ff?mode=detail&view=horizontal&q=Berchem%20&rows=1&page=59&fq%5B%5D=search_s_entity_name:%22Objecten%22.

- Figure 17: *Amsterdam*. Photograph. Europeana (Rijksmuseum). 1880-1920. https://www.europeana.eu/nl/item/90402/RP_F_F19497.

- Figure 18: *Wien, Mariahilfer Straße, Blick zur Stadt*. Photograph. Europeana (Albertina). C. 1890. https://www.europeana.eu/nl/item/15508/Foto2005_166_77.

- Figure 19: *San Diego. 5th st., San Diego, Cal.* Photograph. Europeana (Deutsche Fotothek). 1903. https://www.europeana.eu/nl/item/463/item_T6IBXVO34VDXMPWACLDHKWT62IQUFB62.

- Figure 20: *Korte Nieuwstraat, hoek Melkmarkt, Antwerpen*. Photograph. FelixArchief. 1931. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_6678&dtnr=1224_40&dtrecordid=31973&page=1&pageSize=10&type=copy.

Chapter 5

- Figure 22: *Laken: Maria-Christinastraat*. Photograph. Erfgoedbank Brussel. 1930. https://erfgoedbankbrussel.be/mediabank/detail/38cb5de4-b70e-56ab-0a59-4587cc60214e/media/cde45153-8f39-f99b-e7e2-d54274a4aaa3?mode=detail&view=horizontal&q=laeken%20tram&rows=1&page=.

- Figure 23: *Spårvagn, hästvagn*. Photograph. Europeana (Malmö Museum). 1895-1905. https://www.europeana.eu/nl/item/91672/MM_foto_612814.

- Figure 24: *Zicht op de Steenstraat tussen de Sint-Niklaasstraat en het Simon Stevinplein*. Photograph. Europeana (Stadsarchief Brugge). 1894. https://www.europeana.eu/nl/item/534/377edd52_8121_4ac2_8070_e484627e8a45.

- Figure 25: *Adorf. Markt*. Photograph. Europeana (Deutsche Fotothek). 1907. https://www.europeana.eu/nl/item/437/item_OKFKRYYUJPVTLRFZR6IUKJHTEJQJ6DS5.

- Figure 26: *Sint Joris Gildehuis in Antwerpen*. Photograph. Europeana (Rijksmuseum). 1880-1920. https://www.europeana.eu/nl/item/90402/RP_F_F16285.

- Figure 27: *Grote Markt: de linkervleugel van het Stadhuis, Antwerpen 1928*. Photograph. FelixArchief. 1928. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_3479&dtnr=1224_40&dtrecordid=28872&page=1&pageSize=10&type=copy.

- Figure 28: *East River Bridge N.Y.*. Photograph. Europeana (Swedish National Museum of Science and Technology). 1860-1880. https://www.europeana.eu/nl/item/916118/S_TEK_photo_TEKA0113700.

- Figure 29: *"The Hanson". Hästskjuts i London*. Photograph. Europeana (Swedish National Museum of Science and Technology). 1886. https://www.europeana.eu/nl/item/916118/S_TEK_photo_TEKA0156178.

- Figure 30: *Komotau. Marktplatz*. Photograph. Europeana (Deutsche Fotothek). 1912. https://www.europeana.eu/nl/item/440/item_W27ZVBJUP65WFSVBBKILFTB23FKO7ZHE.

- Figure 31: *Een koets vóór café Heidelberg te Zedelgem*. Photograph. ErfgoedBrugge. N.d. https://erfgoedbrugge.be/p/2300115_19_11987.

- Figure 32: *Leopoldplaats, het ruiterstandbeeld, Antwerpen*. Photograph. FelixArchief. 1910. https://felixarchief.antwerpen.be/detailpagina?invnr=FOTO-OF_4455&dtnr=1224_40&dtrecordid=29812&page=1&pageSize=10&type=copy.

Chapter 6

- Figure 34:

- Epoch 1: Collinet-Guerin. *Rome. Santa Maria del Popolo*. Photograph. Europeana (Institut National d'Histoire de l'Art). 1910. https://www.europeana.eu/nl/item/829/https___bibliotheque_numerique_inha_fr_idurl_1_13755.

- Epoch 10: Moens-Patfoort, Adolf. *Winters zicht op de Groene Rei.* Photograph. Europeana (Stadsarchief Brugge). 1921. https://www.europeana.eu/nl/item/534/b646414b_ea33_4ca8_8373_baeff0f023fd.

- Epoch 20: *Zwickau. Hauptmarkt.* Photograph. Europeana (Deutsche Fotothek). 1915. https://www.europeana.eu/nl/item/437/item_C24VQ6IXSY5RHCMOCDF5QCPH2CFRFKQ2.

- Epoch 30: *Nossen. Markt.* Photograph. Europeana (Deutsche Fotothek). 1913. https://www.europeana.eu/nl/item/437/item_FFWIAXXF6FCPSN47KI42DMQLHRD2RHGD.

- Epoch 40: Böttger, *Georg. Marienplatz mit Hauptwache, München*. Photograph. Europeana (Museum of Arts and Crafts, Hamburg). C. 1865. https://www.europeana.eu/nl/item/2048429/item_JMGH6NIEDVDWPOOBXIFLE4O7A4JT75AV.

- Epoch 50: *Voorstraat. Utrecht*. Photograph. Europeana (Het Utrechts Archief). 1896. https://www.europeana.eu/nl/item/257/https___hetutrechtsarchief_nl_beeld_FF92D46AFBB956EEBCBFCAC9871CB981.

- Epoch 60: Neurdein. *Gezicht op het Provinciaal Hof en het Postkantoor op de Markt*. Photograph. Europeana (Stadsarchief Brugge). 1900. https://www.europeana.eu/nl/item/534/bd2e1f2a_e430_44bc_9d2f_38d4c6436e05.

- Epoch 70: *Cercle Arti et Amicitiae sur le Rokin, Amsterdam.* Photograph. Europeana (Rijksmuseum). 1860. https://www.europeana.eu/nl/item/90402/RP_F_1999_140_5.

- Epoch 80: *Brugge. Vismarkt.* Photograph. Europeana (Katholieke Universiteit Leuven). 1910-1939. https://www.europeana.eu/nl/item/2024903/photography_ProvidedCHO_KU_Leuven_9989494140101488.

- Epoch 90: *Gezicht op de beurs te Rotterdam.* Photograph. Europeana (Rijksmuseum). 1850-1900. https://www.europeana.eu/nl/item/90402/RP_F_F19322 .

- Epoch 100: Marcovici. *Zicht op de zuidzijde van de Burg*. Photograph. Europeana (Stadsarchief Brugge). 1929. https://www.europeana.eu/nl/item/534/392dd3ca_8fa4_4dd8_957e_d474e9e06e24.

- Epoch 110: *Gezicht op de Billingsgate Fish Market te Londen.* Photograph. Europeana (Rijksmuseum). C. 1850-1880. https://www.europeana.eu/nl/item/90402/RP_F_F04676.

- Epoch 120: *Rotterdam. Maasstation. recto, o.* Photograph. Europeana (Het Utrechts Archief). 1900. https://www.europeana.eu/nl/item/257/https___hetutrechtsarchief_nl_beeld_9A2A821FDD59595ABD2747B0B536BF23.

- Epoch 130: De Graeve, Théo. *Gezicht op de Poortersloge in de Academiestraat*. Photograph. Europeana (Stadsarchief Brugge). 1900. https://www.europeana.eu/nl/item/534/d0c060b1_0d8d_4dcf_9e0f_948a52765cb0.

- Epoch 140: *Adorf. Markt*. Photograph. Europeana (Deutsche Fotothek). 1907. https://www.europeana.eu/nl/item/440/item_HJZ6J4YMM3PMHKD23Y7RTE2F7ENDVS6F.

- Epoch 150: *Dom, Utrecht: Ansicht*. Photograph. Europeana (Museum of Architecture at Berlin Institute of Technology). N.d. https://www.europeana.eu/nl/item/08535/item_2SN5QFSAN537SY5AU2LEFXED32XG6VN6.

- Epoch 160: Saarinen. *Horse carriage driver*. Photograph. Europeana (Finnish Heritage Agency). C. 1955. https://www.europeana.eu/nl/item/2021009/_43333935C3426192B2CD32DD3F7379B8.

- Epoch 170: *Champs Elysées, Paris, 1886.* Photograph. Europeana (Swedish National Museum of Science and Technology). 1886. https://www.europeana.eu/nl/item/916118/S_TEK_photo_TEKA0156196.

- Epoch 180: Hersleven, Jacques. *Groentemarkt te Mechelen*. Photograph. Europeana (Koninklijk Instituut voor het Kunstpatrimonium). C. 1930. https://www.europeana.eu/nl/item/2048001/AP_10382387.

- Epoch 190: Duschek, Franz. *Rumänisches Album: Straße in Gent mit dem Belfried bei der Tuchhalle*. Photograph. Europeana (Albertina). C. 1870-1880. https://www.europeana.eu/nl/item/15508/Foto2007_337_44.

- Epoch 200: *Boompjes Rotterdam*. Photograph. Europeana (Rijksmuseum). 1860-1880. https://www.europeana.eu/nl/item/90402/RP_F_F12132.

- Figure 35 (Left to Right):

  - Okänd. *House of Parliament, London, 1886.* Photograph. Euorpeana (Swedish National Museum of Science and Technology). 1886. https://www.europeana.eu/nl/item/916118/S_TEK_photo_TEKA0156191.

  - *Sint Joris Gildehuis in Antwerpen*. Photograph. Europeana (Rijksmuseum). 1880-1920. https://www.europeana.eu/nl/item/90402/RP_F_F16285.

  - Hersleven, Jacques. *Groentemarkt te Mechelen*. Photograph. Europeana (Koninklijk Instituut voor het Kunstpatrimonium). C. 1930. https://www.europeana.eu/nl/item/2048001/AP_10382383.

  - Fançois Beyaert, Henri Joseph; Laureys, Felix ; Naert, Joseph-Jean. *Oostende. Casino Kursaal Exterieur: Zicht op de zijgevel, gezien vanaf de Zeedijk*. Photograph. Europeana (Katholieke Universiteit Leuven). https://www.europeana.eu/nl/item/2024903/photography_ProvidedCHO_KU_Leuven_9989686060101488.

- Figure 36:

  - Epoch 1: Khalilov, Nodir. *A river with boats on it*. Photograph. Unsplash. November 15, 2022. https://unsplash.com/photos/a-river-with-boats-on-it-gUicFkEyMxM.

  - Epoch 10: Yaroshenko, Anastasia. *A cobblestone street in a European city*. Photograph. Unsplash. January 29, 2024. https://unsplash.com/photos/a-cobblestone-street-in-a-european-city-KCsx25rlS4k.

- Epoch 20: Gilly. *Bus and cars on road*. Photograph. Unsplash. October 22, 2018. https://unsplash.com/photos/bus-and-cars-on-road-8vzFINl6zV8.

- Epoch 30: Dobre, Ionut. *A bridge over a body of water with a statue on top*. Photograph. Unsplash. July 7, 2022. https://unsplash.com/photos/a-bridge-over-a-body-of-water-with-a-statue-on-top-nnlVXi1lp3k.

- Epoch 40: Boccia, Peter. *Calm river during daytime*. Photograph. Unsplash. January 7, 2020. https://unsplash.com/photos/calm-river-during-daytime-NZULuaSJQQU.

- Epoch 50: Proposed by Roboflow.

- Epoch 60: Hongbin. *A river running through a city next to tall buildings*. Photograph. Unsplash. July 21, 2023. https://unsplash.com/photos/a-river-running-through-a-city-next-to-tall-buildings-zyw_lt-6r90.

- Epoch 70: Michals, Chris. *A couple of horses pulling a carriage down a street*. Photograph. Unsplash. June 6, 2022. https://unsplash.com/photos/a-couple-of-horses-pulling-a-carriage-down-a-street-cxHA3gphtAY.

- Epoch 80: Katzenberger, Philipp. *A street with buildings on either side*. Photograph. Unsplash. October 17, 2022. https://unsplash.com/photos/a-street-with-buildings-on-either-side-KOFk99anjf4.

- Epoch 90: Wingate, Ed. *Ein Fluss, der durch eine Stadt fließt, neben hohen Gebäuden*. Photograph. Unsplash. August 6, 2024. https://unsplash.com/de/fotos/ein-fluss-der-durch-eine-stadt-fliesst-neben-hohen-gebauden-vbjRA5Z--u0.

- Epoch 100: Proposed by Roboflow.

- Epoch 110: Nikolaieva, Maryna. *A close up of a busy city street*. Photograph. Unsplash. May 8, 2022. https://unsplash.com/photos/a-close-up-of-a-busy-city-street-d4E7VVAJNs4.

- Epoch 120: Danaila, Claudiu. *The Big Ben clock tower towering over the city of London*. Photograph. Unsplash. September 11, 2023. https://unsplash.com/photos/the-big-ben-clock-tower-towering-over-the-city-of-london-IZMShn693RI.

- Epoch 130: Paulin, Louis. *People walking on park near white concrete building during daytime*. Photograph. Unsplash. February 9, 2020. https://unsplash.com/photos/people-walking-on-park-near-white-concrete-building-during-daytime-qWlRwhVkHtU.

- Epoch 140: Rouiller, Alain. *A group of people sitting on a bench next to a river*. Photograph. Unsplash. March 25, 2024. https://unsplash.com/photos/a-group-of-people-sitting-on-a-bench-next-to-a-river-Zi7jbQY9nyU.

- Epoch 150: Proposed by Roboflow.

- Epoch 160: Gio. *People walking on street during nighttime*. Photograph. Unsplash. May 22, 2020. https://unsplash.com/photos/people-walking-on-street-during-nighttime-8tzxy0Gxf38.

- Epoch 170: Komissarov, Alexey. *A trolley car is traveling down the street*. Photograph. Unsplash. July 1, 2024. https://unsplash.com/photos/a-trolley-car-is-traveling-down-the-street-pFHX6o4hsG4.

- Epoch 180: Sachadonig, Vali. *Yellow and white bus in the middle of white buildings during daytime*. Photograph. Unsplash. October 8, 2017. https://unsplash.com/photos/yellow-and-white-bus-in-the-middle-of-white-buildings-during-daytime-hVzHCspuUM4.

- Epoch 190: Toader, Andrei. *Vehicle beside buildings during daytime*. Photograph. Unsplash. October 3, 2019. https://unsplash.com/photos/vehicle-beside-buildings-during-daytime-QHhNbQq7-ps.

- Epoch 200: Peeters, Florian. *A city street with cars parked on both sides*. Photograph. Unsplash. June 19, 2023. https://unsplash.com/photos/a-city-street-with-cars-parked-on-both-sides-gBWhIm1JPB4.

- Figure 37 (Left to Right):

  - Wei, Justin. *A woman with a backpack and a crown on her head*. Photograph. Unsplash. June 11, 2024. https://unsplash.com/photos/a-woman-with-a-backpack-and-a-crown-on-her-head-9CwctZwGtcY.

- Nikolaieva, Maryna. *People standing near brown and red concrete building during daytime*. Photograph. Unsplash. June 22, 2021. https://unsplash.com/photos/people-standing-near-brown-and-red-concrete-building-during-daytime-4P88cZe-dzE.

- Heymans, Peter. *White and black train on rail road near green trees during daytime*. Photograph. Unsplash. August 27, 2021. https://unsplash.com/photos/white-and-black-train-on-rail-road-near-green-trees-during-daytime-O-CoALJbEMQ.

- Santos, Elio. *Crowd rallying outside buildings*. Photograph. Unsplash. May 7, 2019. https://unsplash.com/photos/crowd-rallying-outside-buildings-4GV7AQn6xmQ.

- Figure 39:

  - *Stadhuis Antwerpen.* Photograph. FelixArchief. C. 1870-1882. https://felixarchief.antwerpen.be/detailpagina?invnr=934_21909&dtnr=1224_40&dtrecordid=110084&page=1&pageSize=10&type=copy.

Colab Notebook

Cuykens, Vastert. 2025. *The Sound of the City. Google Colab Notebook Containing All Used Code.* Google Colab. https://colab.research.google.com/drive/196m5kAn-IMC0fBAVT4ZK7Wp85tkoHzas?usp=sharing.

Literature

"Add Fasterrcnn Improved Weights #5763." GitHub, accessed 10 April 2025, https://github.com/pytorch/vision/pull/5763.

"Add Retinanet Improved Weights #5756." GitHub, accessed 10 April 2025, https://github.com/pytorch/vision/pull/5756.

Ali, Momina Liaqat, and Zhou Zhang. "The Yolo Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection." *Computers* 13, 12 (2024). https://doi.org/10.3390/computers13120336.

Andres, Alain, Aitor Martinez-Seras, Ibai Laña, and Javier Del Ser. "On the Black-Box Explainability of Object Detection Models for Safe and Trustworthy Industrial Applications." *Results in Engineering* 24 (2024). https://doi.org/10.1016/j.rineng.2024.103498.

"Artificial Intelligence and Copyright: Use of Generative Ai Tools to Develop New Content." European Commission, 2024, accessed 4 April 2025, https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_en.

"Automatically Label Image Data." Roboflow, accessed 14 November 2024, https://roboflow.com/auto-label.

Bai, Jianwei. "Ancient Chinese Painting Style Transfer Based on Cyclegan." *Applied and Computational Engineering* 51, no. 1 (2024): 129-36. https://doi.org/10.54254/2755-2721/51/20241192.

Bender, Ana, Manuela Guerreiro, Dora Agapito, Bernardete Dias Sequeira, and Júlio Mendes. "Sensory Experiences in Heritage Contexts: A Qualitative Approach.". *European Journal of Tourism Research* 36 (2024). https://doi.org/10.54055/ejtr.v36i.3060.

Bengamra, Siwar, Olfa Mzoughi, André Bigand, and Ezzeddine Zagrouba. "A Comprehensive Survey on Object Detection in Visual Art: Taxonomy and Challenge." *Multimedia Tools and Applications* 83, no. 5 (2024): 14637–70. https://doi.org/10.1007/s11042-023-15968-9.

Berardi, Umberto, Gino Iannace, and Luigi Maffei. "Virtual Reconstruction of the Historical Acoustics of the Odeon of Pompeii." *Journal of Cultural Heritage* 19 (2016): 555-66. https://doi.org/10.1016/j.culher.2015.12.004.

"Guide to Generative Adversarial Networks (Gans) in 2025." viso.ai, Updated 1 October 2024, 2024, accessed 24 December 2024, https://viso.ai/deep-learning/generative-adversarial-networks-gan/.

"Bretez." accessed 10 April 2025, https://sites.google.com/site/louisbretez/accueil.

Briatore, Samuele. "Immaginare I Suoni. Ricostruzione Del Paesaggio Sonoro Della Festa Barocca." *Arti dello Spettacolo / Performing Arts* 3, no. 3 (2017).

Capurro, Carlotta, and Gertjan Plets. "Europeana, Edm, and the Europeanisation of Cultural Heritage Institutions." *Digital Culture & Society* 6, no. 2 (2020): 163-90. https://doi.org/10.14361/dcs-2020-0209.

Carneiro, Tiago, Raul Victor Medeiros Da NóBrega, Thiago Nepomuceno, Gui-Bin Bian, Victor Hugo C. De Albuquerque, and Pedro Pedrosa Rebouças Filho. "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications." *IEEE Access* 6 (2018): 61677-85. https://doi.org/10.1109/ACCESS.2018.2874767.

Chen, Yen-Chia, Hiroki Shibata, Lieu-Hen Chen, and Yasufumi Takama. "Synthesis of Comic-Style Portraits Using Combination of Cyclegan and Pix2pix." *Journal of Advanced Computational Intelligence and Intelligent Informatics* 28, no. 5 (2024): 1085-94. https://doi.org/10.20965/jaciii.2024.p1085.

Crasto, Nieves. "Class Imbalance in Object Detection: An Experimental Diagnosis and Study of Mitigation Strategies." *arXiv (Cornell University)* (2024). https://doi.org/10.48550/arXiv.2403.07113.

"Generative Adversarial Networks." Medium, Updated 30 October 2023, 2023, accessed 24 December 2024, https://medium.com/@marcodelpra/generative-adversarial-networks-dba10e1b4424.

"Soundscape Archaeology: A Visit to Paris in the Mid 18th/Early 19th Century." Fondation Napoléon, 2020, accessed 10 April 2025, https://www.napoleon.org/en/history-of-the-two-empires/videos/soundscape-archaeology-a-visit-to-paris-in-the-mid-18th-early-19th-century/.

Doan, Anh-Dzung, Bach Long Nguyen, Surabhi Gupta, Ian Reid, Markus Wagner, and Tat-Jun Chin. "Assessing Domain Gap for Continual Domain Adaptation in Object Detection." *Computer Vision and Image Understanding* 238 (2024). https://doi.org/10.1016/j.cviu.2023.103885.

"Erfgoed Brugge." Erfgoed Brugge, accessed 10 October 2024, https://erfgoedbrugge.be/.

"Erfgoedbank Brussel." Erfgoedbank Brussel, accessed 10 October 2024, https://erfgoedbankbrussel.be/.

"Europeana." Europeana, accessed 10 October 2024, https://www.europeana.eu/nl.

"Explore Ultralytics Yolov8." Ultralytics, accessed 10 April 2025, https://docs.ultralytics.com/models/yolov8/.

"Fasterrcnn_Resnet50_Fpn_V2." PyTorch, accessed 10 January 2025, https://pytorch.org/vision/main/models/generated/torchvision.models.detection.fasterrcnn_resnet50_fpn_v2.html#torchvision.models.detection.fasterrcnn_resnet50_fpn_v2.

"Felixarchief." Stad Antwerpen, accessed 10 October 2024, https://felixarchief.antwerpen.be/.

"Fine-Tuning." Ultralytics, accessed 10 April 2025, https://www.ultralytics.com/glossary/fine-tuning.

Fiorucci, Marco , Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. "Machine Learning for Cultural Heritage: A Survey." *Pattern Recognition Letters* 133 (2020): 102-08. https://doi.org/10.1016/j.patrec.2020.02.017.

Firat, Hasan Baran. "Acoustics as Tangible Heritage: Re-Embodying the Sensory Heritage in the Boundless Reign of Sight." *Preservation, Digital Technology & Culture* 50, no. 1 (2021): 3-14. https://doi.org/10.1515/pdtc-2020-0028.

Firat, Hasan Baran, and Luigi Maffei. "A Methodology for the Historically Informed Soundscape." Inter-.Noise 2020, e-congress, 2020.
Firat, Hasan Baran, Luigi Maffei, and Massimiliano Masullo. "Digital Humanities in the Historical Soundscape Research: Sounds of 18th Century Naples." The Acoustics of Ancient Theatres, Verona, Italy, 2022.
"Fotografisch Geheugen: De Reportages Van Fotopersbureau De Boer." VeleHanden, accessed 22 April 2025, https://velehanden.nl/projecten/bekijk/details/project/ranh_tagselection_deboer.
Galvano, Francesco. "The Triad of Senses, Emotions, and Memory: Dynamic Interactions and Multidisciplinary Implications." (2015): 1-12.
"Launch: Synthetic Image Generation with Dall-E and Gpt-4 Vision." Roboflow, 2023, accessed 3 May 2025, https://blog.roboflow.com/synthetic-data-dall-e-roboflow/.
Ultralytics Yolov8.
Goodfellow, Ian, Jean Pouget-Abadie, Mehd Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." *arXiv (Cornell University)* (2014): 1-9. https://doi.org/10.48550/arxiv.1406.2661.
"Google Colaboratory." Google, accessed 14 November 2024, https://colab.google/#:~:text=Google%20Colaboratory,%2C%20data%20science%2C%20and%20education.
Graybill, Rhiannon. "'Hear and Give Ear!': The Soundscape of Jeremiah." *Journal for the Study of the Old Testament* 40, no. 4 (2016): 467-90. https://doi.org/10.1177/0309089216628414.
Hatir, Ergün, Korkanç Mustafa, Andreas Schacner, and İsmail İnce. "The Deep Learning Method Applied to the Detection and Mapping of Stone Deterioration in Open-Air Sanctuaries of the Hittite Period in Anatolia." *Journal of Cultural Heritage* 51 (2021): 37-49. https://doi.org/10.1016/j.culher.2021.07.004.
"Hearing the Past: Reconstructing the Aural Heritage of Antwerp in the 19th Century." University of Antwerp, accessed 10 April 2025, https://www.uantwerpen.be/en/projects/hearing-the-past/about/.
Hindarto, Djarot. "Revolution in Image Data Collection: Cyclegan as a Dataset Generator." *Sinkron : Jurnal Dan Penelitian Teknik Informatika* 8, no. 1 (2024): 444-54. https://doi.org/10.33395/sinkron.v9i1.13211.
Howes, David. "Introduction to Sensory Museology." *The Senses and Society* 9, no. 3 (2014): 259-67. https://doi.org/10.2752/174589314X14023847039917.
Ibrahim, Bekkouch Imad Eddine, Victoria Eyharabide, Valérie Le Page, and Frédéric Billiet. "Few-Shot Object Detection: Application to Medieval Musicological Studies." *Journal of Imaging* 8, no. 18 (2022). https://doi.org/10.3390/jimaging8020018.
Jabir, Brahim, Noureddine Falir, and Khalid Rahmani. "Accuracy and Efficiency Comparison of Object Detection Open-Source Models." *International Journal of Online and Biomedical Engineering* 15, no. 5 (2021): 165-83. https://doi.org/10.3991/ijoe.v17i05.21833.
Kadish, David, Sebastian Risi, and Anders Sundess Løvlie. "Improving Object Detection in Art Images Using Only Style Transfer." *arXiv (Cornell University)* (2021). https://doi.org/10.48550/arxiv.2102.06529.
Kang, Jian, Francesco Aletta, Truls T. Gjestland, Lex A. Brown, Dick Botteldooren, Brigitte Schulte-Fortkamp, Peter Lercher*, et al.* "Ten Questions on the Soundscapes of the Built Environment." *Building and Environment* 108 (2016): 284-94. https://doi.org/10.1016/j.buildenv.2016.08.011.
Karbouj, Bsher, Garabet A. Topalian-Rivas, and Jörg Krüger. "Comparative Performance Evaluation of One-Stage and Two-Stage Object Detectors for Screw Head Detection and Classification in Disassembly Processes." *Procedia CIRP* 122 (2024): 527-32. https://doi.org/10.1016/j.procir.2024.01.077.
Kim, Yongho, Chanjong Im, and Thomas Mandl. "Object Detection in Historical Images: Transfer Learning and Pseudo Labelling." *ACM Journal on Computing and Cultural Heritage* (2024): 1-15. https://doi.org/10.1145/3699963.
"Retinanet: Single-Stage Object Detector with Accuracy Focus." viso.ai, 2024, accessed 3 April 2025, https://viso.ai/deep-learning/retinanet/.

Kulkarni, Uday, Meena S.M., Sunil V. Gurlahosur, and Uma Mudengudi. "Classification of Cultural Heritage Sites Using Transfer Learning." IEEE Fifth International Conference on Multimedia Big Data (BigMM), Signapore, 2019.
Lee, Benjamin Charles Germain. "The “Collections as Ml Data” Checklist for Machine Learning and Cultural Heritage." *Journal of the Association for Information Science and Technology* 76, no. 2 (2023): 375-96. https://doi.org/10.1002/asi.24765.
Leon-Gomez, Eder Arley, Andrés Marino Álvarez-Meza, and German Castellanos-Dominguez. "Cross-Dataset Data Augmentation Using Umap for Deep Learning-Based Wind Speed Prediction." *Computers* 14, no. 4 (2025): 123-43. https://doi.org/10.3390/computers14040123.
"Library of Congress." Library of Congress, accessed 10 October 2024, https://www.loc.gov/.
Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal Loss for Dense Object Detection." *arXiv (Cornell University)* (2017). https://doi.org/10.48550/arXiv.1708.02002.
Luo, Dan, Lieve Doucé, and Karin Nys. "Multisensory Museum Experience: An Integrative View and Future Research Directions." *Museum Management and Curatorship* (2024): 1-28. https://doi.org/10.1080/09647775.2024.2357071.
Luo, Vinson, Michael Straka, and Lucy Li. "Historical and Modern Image-to-Image Translation with Generative Adversarial Networks." (2017): 1-8.
Maiwald, Ferdinand. "Generation of a Benchmark Dataset Using Historical Photographs for Automated Evaluation of Different Feature Matching Methods." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, no. 2 (2019): 87-94. https://doi.org/10.5194/isprs-archives-XLII-2-W13-87-2019.
Malcolm-Davies, Jane. "Structuring Reconstructions: Recognising the Advantages of Interdisciplinary Data in Methodical Research." *Heritage Science* 11, no. 1 (2023): 1-16. https://doi.org/10.1186/s40494-023-00982-9.
"Ai ‘Opt-Outs’: Should Cultural Heritage Institutions (Dis)Allow the Mining of Cultural Heritage Data?" Europeana Pro, 2024, accessed 4 April 2025, https://pro.europeana.eu/post/ai-opt-outs-should-cultural-heritage-institutions-dis-allow-the-mining-of-cultural-heritage-data.
Monna, Fabrice, Tanguy Rolland, Anthony Denaire, Nicolas Navarro, Ludovic Granjon, Rémi Barbé, and Carmela Chateau-Smith. "Deep Learning to Detect Built Cultural Heritage from Satellite Imagery. - Spatial Distribution and Size of Vernacular Houses in Sumba, Indonesia." *Journal of Cultural Heritage* 52 (2021): 171-83. https://doi.org/10.1016/j.culher.2021.10.004.
Mumuni, Alhassan, and Fuseini Mumuni. "Data Augmentation: A Comprehensive Survey of Modern Approaches." *Array* 16 (2022). https://doi.org/10.1016/j.array.2022.100258.
Müske, Johannes. "Constructing Sonic Heritage: The Accumulation of Knowledge in the Context of Sound Archives." *The Journal of Ethnology and Folkloristics* 4, no. 1 (2011): 37-47.
"What Is Yolo? The Ultimate Guide [2025]." Roboflow, 2025, accessed 2 April 2025, https://blog.roboflow.com/guide-to-yolo-models/#yolov12.
Padilla, Rafael, Sergio L. Netto, and Eduardo A. B. da Silva. "A Survey on Performance Metrics for Object-Detection Algorithms." In *Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (Iwssip)*, edited by Anselmo C. Paiva, Aura Conci, Geraldo Braz Jr., João Dallyson S Almeida and Leandro A. F. Fernandes, 237-42: Institute of Computing at Fluminense Federal University (IC-UFF), 2020.
Pai, Suraj, Ibrahim Hadzic, Chinmay Rao, Ivan Zhovannik, Andre Dekker, Alberto Traverso, Stylianos Asteriadis, and Enrique Hortal. "Frequency-Domain-Based Structure Losses for Cyclegan-Based Cone-Beam Computed Tomography Translation." *Sensors* 23, no. 3 (2023). https://doi.org/10.3390/s23031089.
Pardoen, Mylène. "Projet Bretez: Une Pincée De Son Dans L’histoire.". *Digital Studies/Le champ numérique* 9, no. 1 (2019): 1-17. https://doi.org/10.16995/dscn.350.
"Restoring the Historical Sound of Paris." Hypotheses, 2016, accessed 10 April 2025, https://sms.hypotheses.org/8560.
Park, Jaihyun, David K. Han, and Hanseok Ko. "Adaptive Weighted Multi-Discriminator Cyclegan for Underwater Image Enhancement." *Journal of Marine Science and Engineering* 7, no. 7 (2019). https://doi.org/10.3390/jmse7070200.

Patra, Aditya, and Rae Crandall. "Machine Learning for Visually Impaired: Benchmarking Object Detection Models." *Journal of Student Research* 13, no. 2 (2024): 1-14. https://doi.org/10.47611/jsrhs.v13i2.6630.

Pietroni, Eva. "Mapping the Soundscape in Communicative Forms for Cultural Heritage: Between Realism and Symbolism." *Heritage* 4, no. 4 (2021): 4495-523. https://doi.org/10.3390/heritage4040248.

Poudel, Anil. "Face Recognition on Historical Photographs." Uppsala University, 2021.

"Pseudo Labeling: Leveraging the Power of Self-Supervision in Machine Learning." Medium, Updated 1 February 2024, 2024, accessed 27 December 2024, https://medium.com/@data-overload/pseudo-labeling-leveraging-the-power-of-self-supervision-in-machine-learning-d8192e918d65.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection." *arXiv (Cornell University)* (2015). https://doi.org/10.48550/arXiv.1506.02640.

"Regulation (Eu) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (Ec) No 300/2008, (Eu) No 167/2013, (Eu) No 168/2013, (Eu) 2018/858, (Eu) 2018/1139 and (Eu) 2019/2144 and Directives 2014/90/Eu, (Eu) 2016/797 and (Eu) 2020/1828 (Artificial Intelligence Act)." edited by European Union and European Council, 2024. http://data.europa.eu/eli/reg/2024/1689/oj.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks." *arXiv (Cornell University)* (2015). https://doi.org/10.48550/arXiv.1506.01497.

"Retinanet_Resnet50_Fpn_V2." PyTorch, accessed 10 January 2025, https://pytorch.org/vision/0.20/models/generated/torchvision.models.detection.retinanet_resnet50_fpn_v2.html.

"Roboflow." Roboflow, accessed 14 November 2024, https://roboflow.com/.

"Train-Yolov8-Object-Detection-on-Custom-Dataset.Ipynb." Google Colab, accessed 24 January 2025, https://colab.research.google.com/github/roboflow-ai/notebooks/blob/main/notebooks/train-yolov8-object-detection-on-custom-dataset.ipynb.

Sapkota, Ranjan, and Manoj Karkee. "Generative Ai in Agriculture: Creating Image Datasets Using Dall.E's Advanced Large Language Model Capabilities." *arXiv (Cornell University)* (2023). https://doi.org/10.48550/arXiv.2307.08789.

Schafer, R. Murray. "Open Ears." *Soundscape: The Journal of Acoustic Ecology* 4, no. 2 (2003): 14-18.

———. *The Soundscape: Our Sonic Environment and the Tuning of the World.* New York: Alfred Knopf, Inc., 1977.

Sender Contell, Marina, Ana Planells, Ricardo Perelló Roso, Jaume Segura Garcia, and Alicia Giménez. "Virtual Acoustic Reconstruction of a Lost Church: Application to an Order of Saint Jerome Monastery in Alzira, Spain." *Journal of Building Performance Simulation* 11, no. 3 (2018): 369-90. https://doi.org/10.1080/19401493.2017.1340975.

Shin, Jeongmin, and Hyeryung Jang. "Data Augmentation Techniques Using Text-to-Image Diffusion Models for Enhanced Data Diversity." 2024 15th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, IEEE, 2024.

Shorten, Connor, and Taghi M Khoshgoftaar. "A Survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6, no. 1 (2019). https://doi.org/10.1186/s40537-019-0197-0.

"Smell Heritage – Sensory Mining." accessed 30 March 2025, https://odeuropa.eu/.

Sohan, Mupparaju, Thotakura SaiRam, and Ch. Venkata RamiReddy. "A Review on Yolov8 and Its Advancements." In *Data Intelligence and Cognitive Informatics. Proceedings of Icdici 2023.*, edited by I. Jeena Jacob, Selwyn Piramuthu and Przemyslaw Falkowski-Gilski. Algorithms for Intelligent Systems, 529-45. Singapore: Springer Singapore, 2024.

"Sonic Heritage - Exploring the Sounds of the World's Most Famous Sights." Cities and Memory, Oxford, accessed 31 March 2025, https://citiesandmemory.com/heritage/.

"The Soundscape of Istanbul Collection - About Collection." Koç University, accessed 31 March 2025, https://librarydigitalcollections.ku.edu.tr/en/collection/soundscape-of-istanbul/.

Terven, Juan, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. "A Comprehensive Review of Yolo Architectures in Computer Vision: From Yolov1 to Yolov8 and Yolo-Nas." *Machine Learning and Knowledge Extraction* 5, no. 4 (2023): 1680-716. https://doi.org/10.3390/make5040083.

Tong, Zhaoxiang. "Exploring the Impact of Hyperparameters on the Generation Quality of Cyclegan." *Transactions on Computer Science and Intelligent Systems Research* 5 (2024): 265-71. https://doi.org/10.62051/01m93a63.

"Transfer Learning with Frozen Layers in Yolov5." Ultralytics, accessed 10 April 2025, https://docs.ultralytics.com/yolov5/tutorials/transfer_learning_with_frozen_layers/.

Ulmer, Maximilian, Leonard Klüpfel, Maximilian Durner, and Rudolph Triebel. "How Important Are Data Augmentations to Close the Domain Gap for Object Detection in Orbit?". *arXiv (Cornell University)* (2024). https://doi.org/10.48550/arXiv.2410.15766.

"Ultralytics Yolo Hyperparameter Tuning Guide." Ultralytics, accessed 24 April 2025, https://docs.ultralytics.com/guides/hyperparameter-tuning/.

UNESCO. "39 C/49 the Importance of Sound in Today's World: Promoting Best Practices." General Conference, 39th session, Paris 2017.

———. *Basic Texts of the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage.* 2022 ed. Paris: France, 2022.

"Unsplash." Unsplash, accessed 10 October 2024, https://unsplash.com/.

"All You Need to Know About Explainable Ai (Xai)." Ultralytics, 2024, accessed 17 April 2025, https://www.ultralytics.com/blog/all-you-need-to-know-about-explainable-ai.

Wei, Jian, Qinzhao Wang, and Zixu Zhao. "Yolo-G: Improved Yolo for Cross-Domain Object Detection." *PLoS ONE* 18, no. 9 (2023). https://doi.org/10.1371/journal.pone.0291241.

Wevers, Melvin. "Scene Detection in De Boer Historical Photo Collection." In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, edited by Ana Paula Rocha, Luc Steels and Jaap Van den Herik, 301-610. Cham: Springer, 2021.

"What Is Cities and Memory?" Cities and Memory, Oxford, accessed 31 March 2025, https://citiesandmemory.com/what-is-cities-and-memory-about/.

Yelmi, Pinar. "Protecting Contemporary Cultural Soundscapes as Intangible Cultural Heritage: Sounds of Istanbul." *International Journal of Heritage Studies* 22, no. 4 (2016): 302-3011. https://doi.org/10.1080/13527258.2016.1138237.

Yelmi, Pınar, Hüseyin Kuşcu, and Asım Evren Yantaç. "Towards a Sustainable Crowdsourced Sound Heritage Archive by Public Participation: The Soundsslike Project." 9th Nordic Conference on Human-Computer Interaction, Gothenburg, Sweden, Association for Computing Machinery, New York, 2016.

Yin, Yuwei, Jean Kaddour, Xiang Zhang, Yixin Nie, Zhenguang Liu, Lingpeng Kong, and Qi Liu. "Ttida: Controllable Generative Data Augmentation Via Text-to-Text and Text-to-Image Models." *arXiv (Cornell University)* (2023). https://doi.org/10.48550/arXiv.2304.08821.

"Pytorch-Cyclegan-and-Pix2pix." GitHub, Updated 22 March 2024, accessed 4 January 2025, https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix/tree/master.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." *arXiv (Cornell University)* (2017). https://doi.org/10.48550/arxiv.1703.10593.

Zinnen, Mathias, Prathmesh Madhu, Ronak Kosti, Peter Bellz, Andreas Maier, and Vincent Christlein. "Odor: The Icpr2022 Odeuropa Challenge on Olfactory Object Recognition." 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022.

Annexes

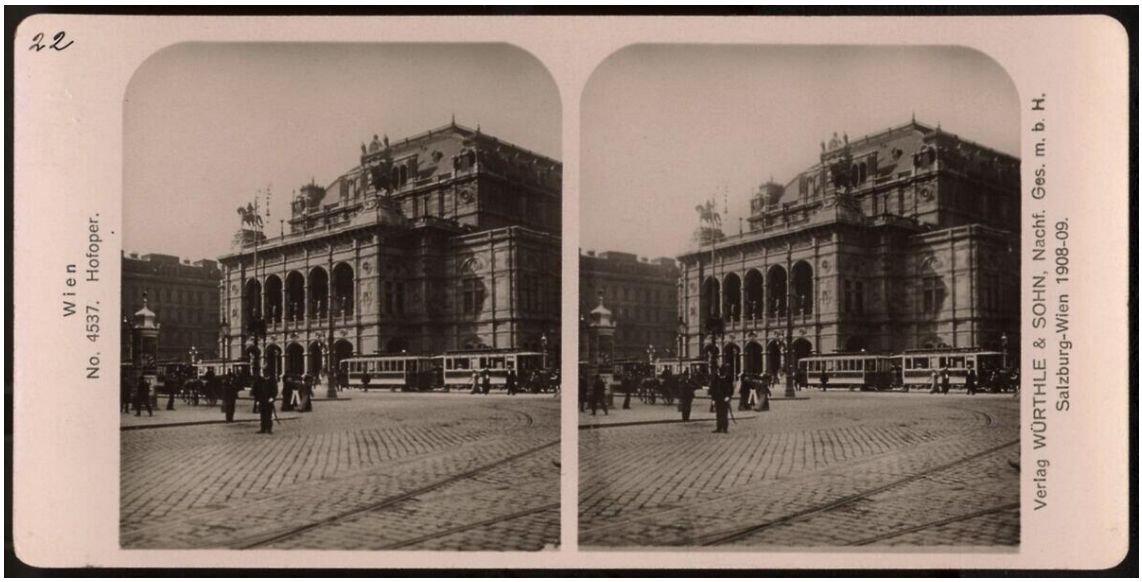
Annex 1: Original photograph Figure 5

Annex 2: Original photograph Figure 6

Annex 3: Original photograph Figure 7

Annex 4: Original photograph Figure 8

Annex 5: Original photograph Figure 9

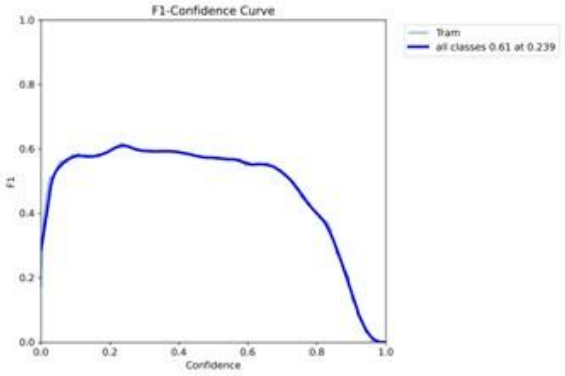


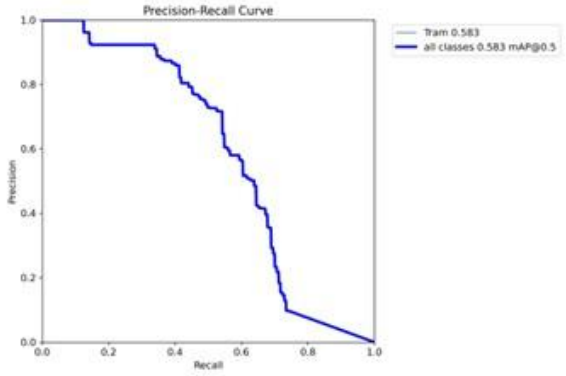
Annex 6: Original photograph Figure 10

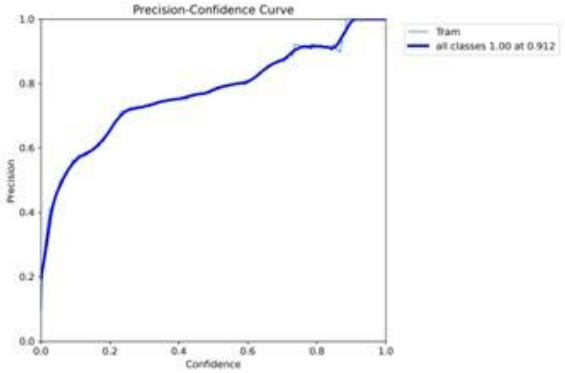
Annex 7: Original photograph Figure 11

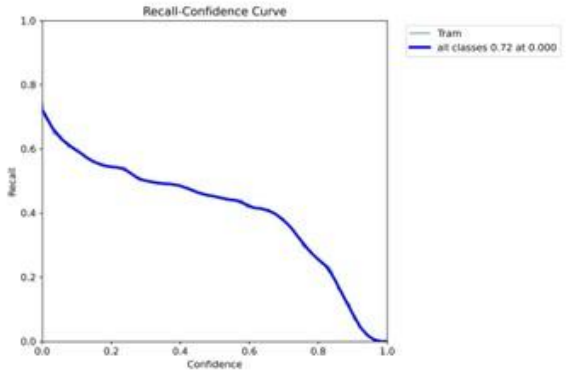
Annex 8: Original photograph Figure 12

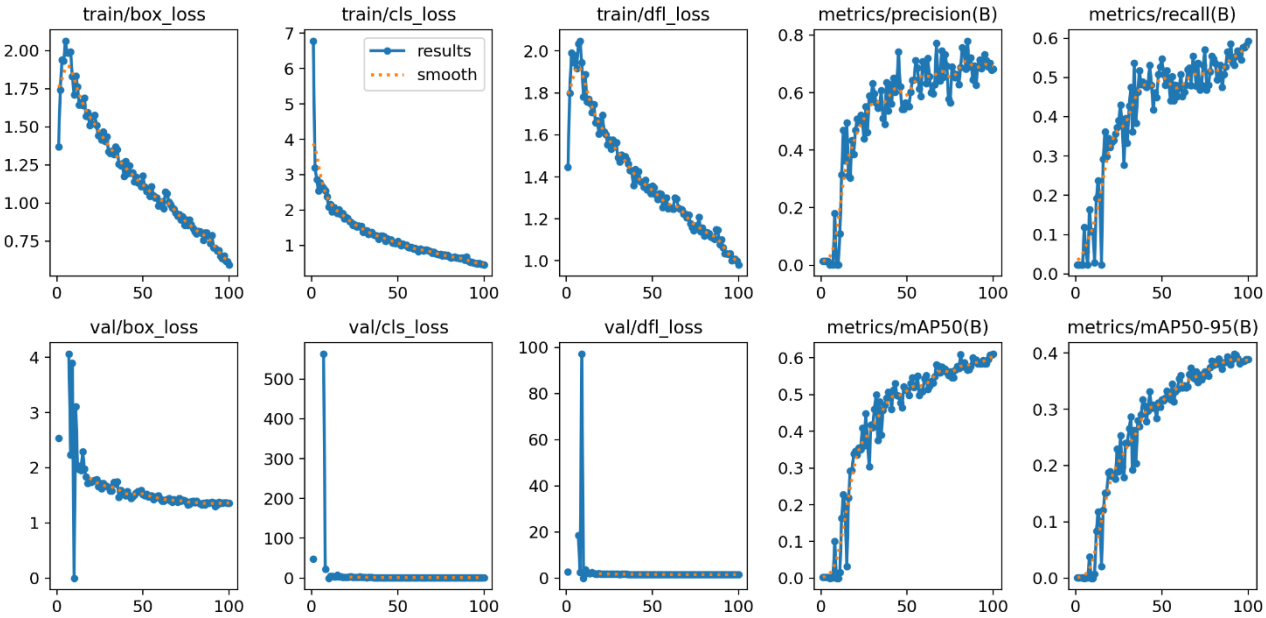
Annex 9: D0-Base Detailed Results

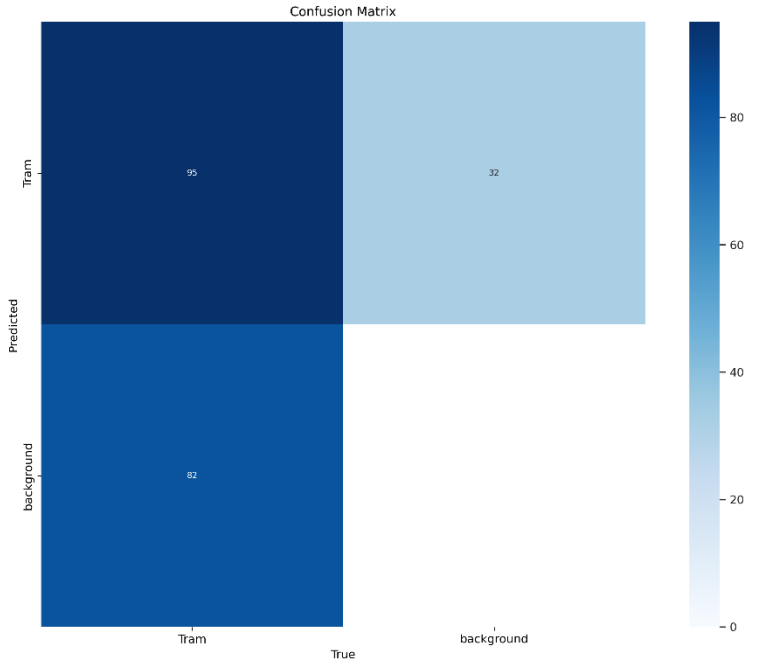


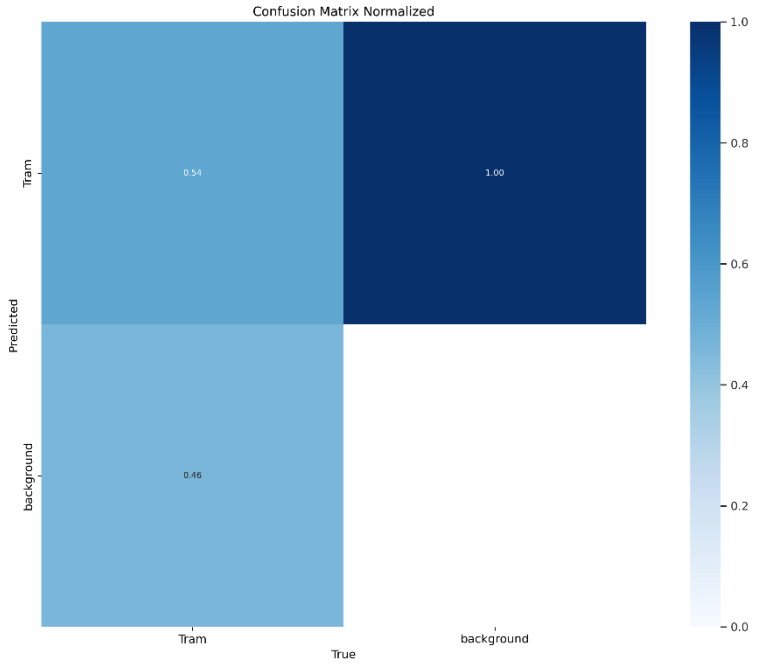




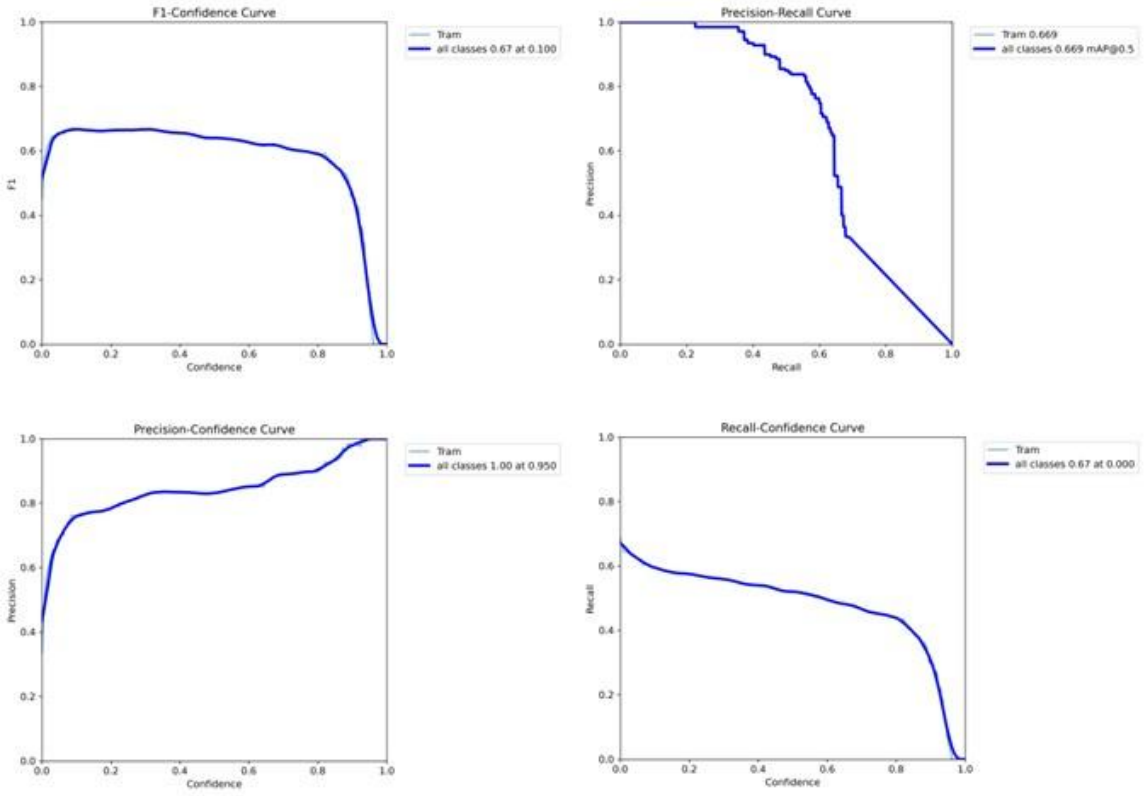


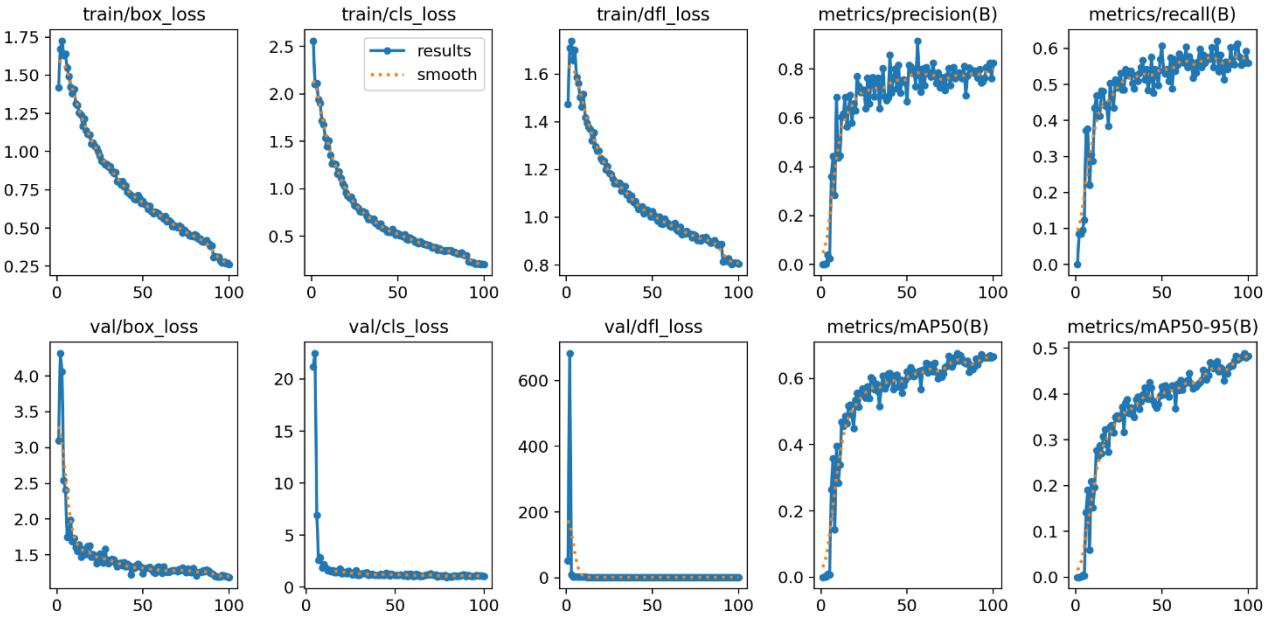


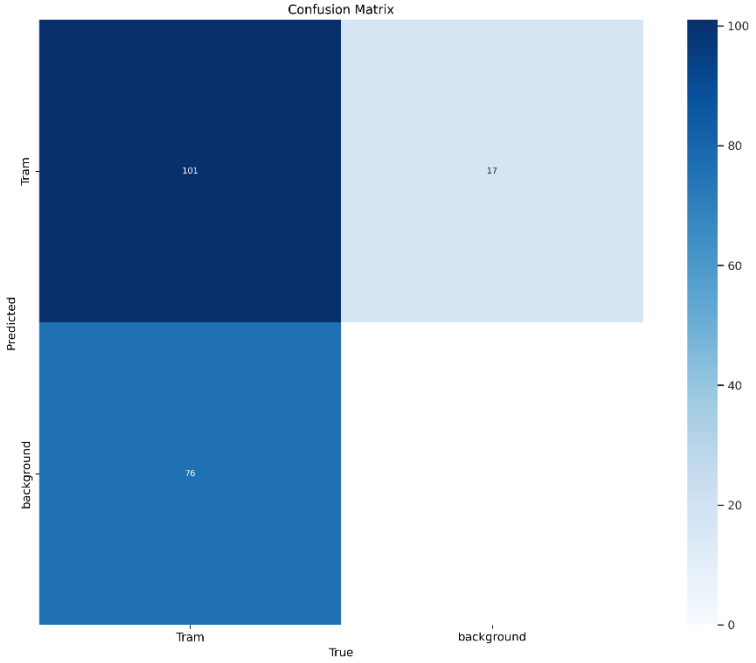

Confusion Matrix
Tram
background
Predicted
95
32
82
Tram
background
True
80
60
40
20
0

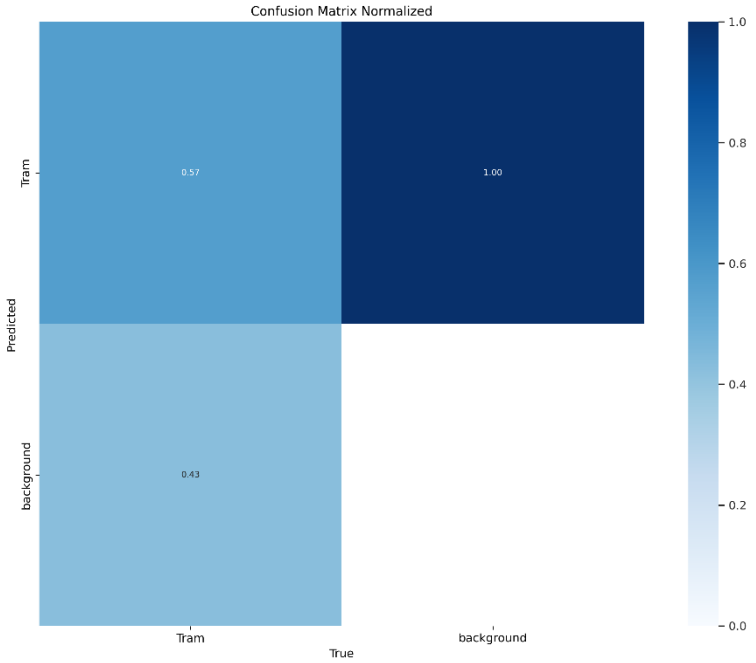

Confusion Matrix Normalized
Tram
background
Predicted
0.54
1.00
0.46
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

Annex 10: D1-BRT Detailed Results

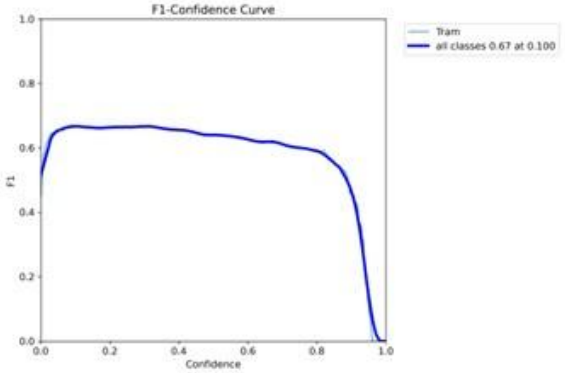


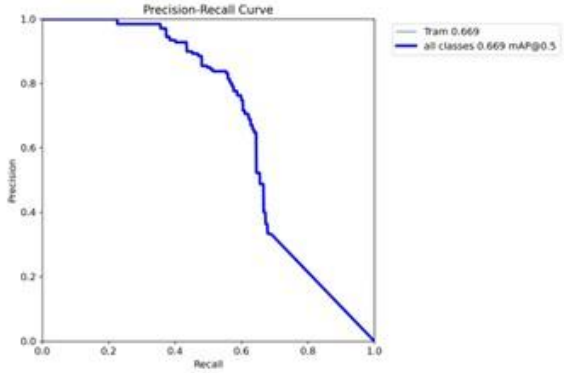


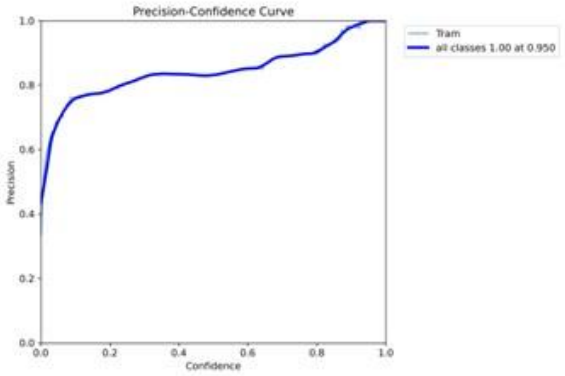

Confusion Matrix
Tram
background
Predicted
101
17
76
Tram
background
True
100
80
60
40
20
0

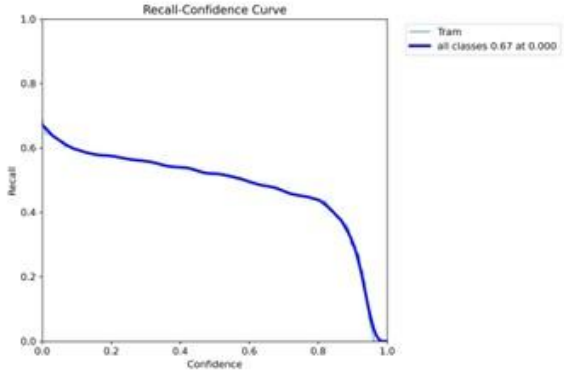

Confusion Matrix Normalized
Tram
background
Predicted
0.57
1.00
0.43
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

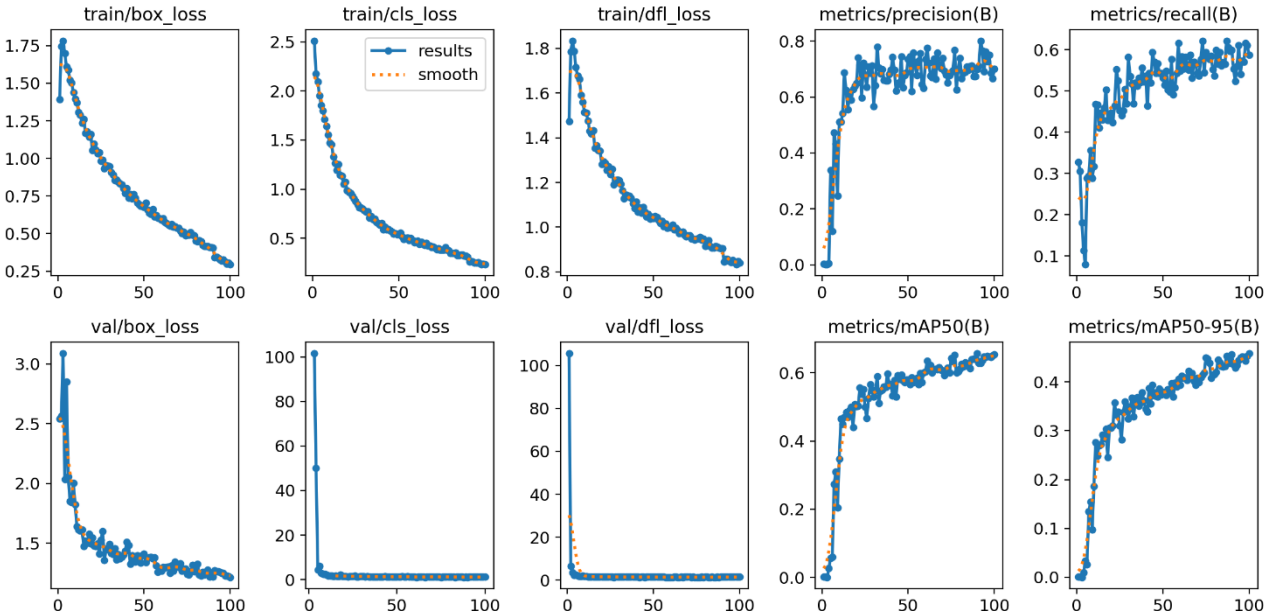
Annex 11: D2-EXP Detailed Results

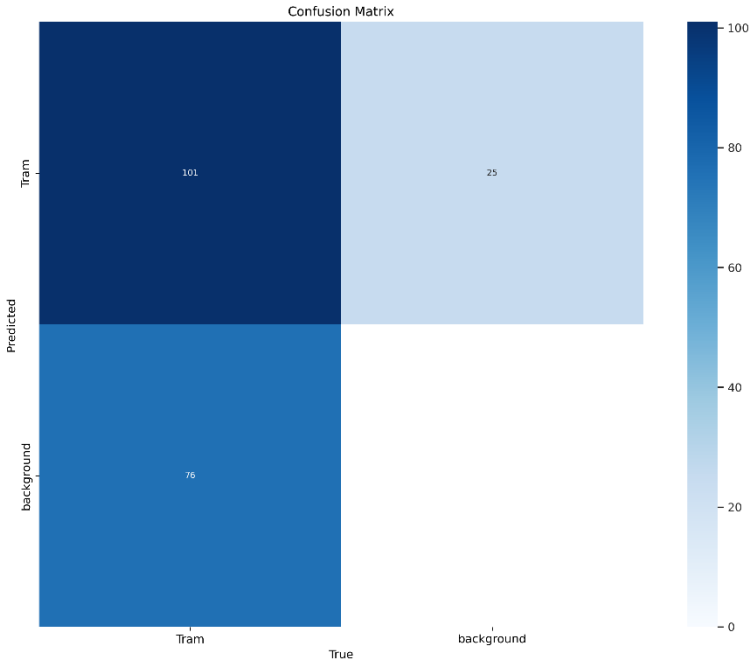


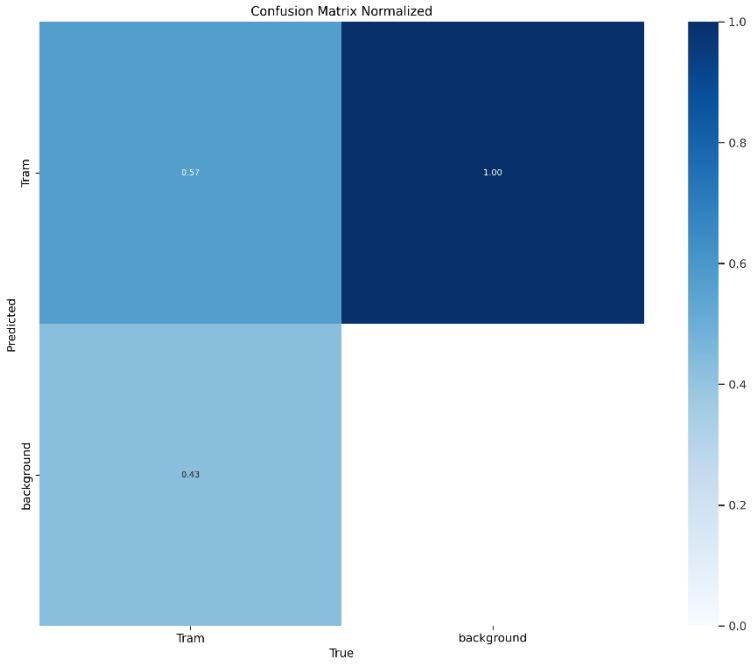




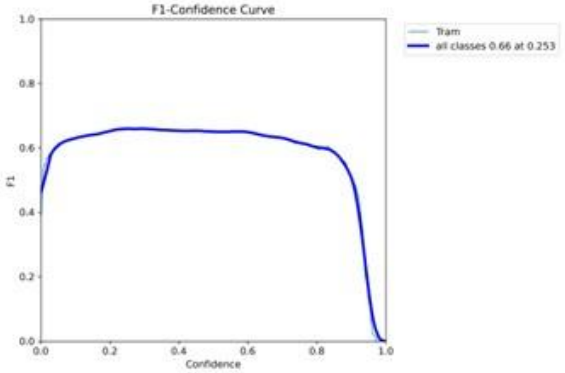


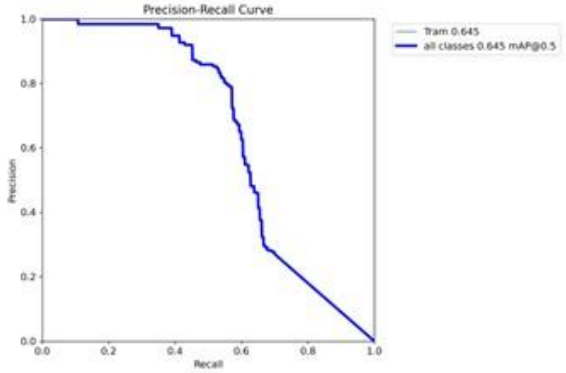


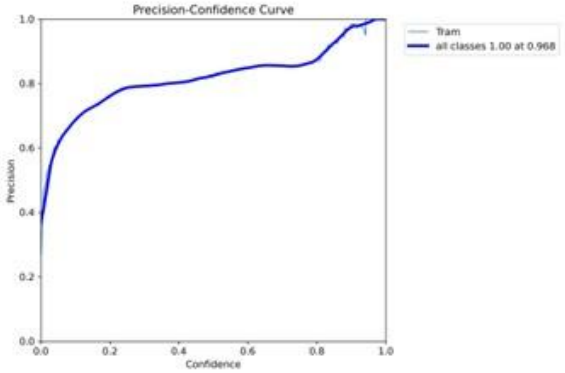

Confusion Matrix
Tram
background
Predicted
101
25
76
Tram
background
True
100
80
60
40
20
0

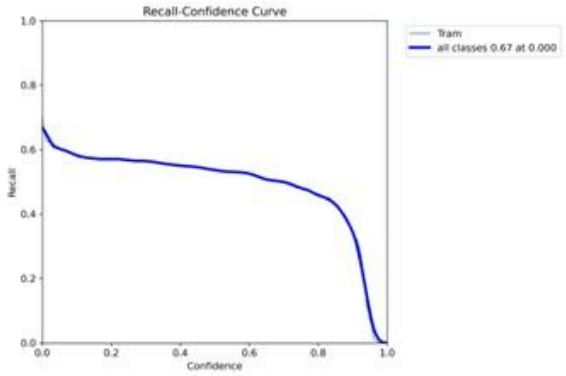

Confusion Matrix Normalized
Tram
background
Predicted
0.57
1.00
0.43
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

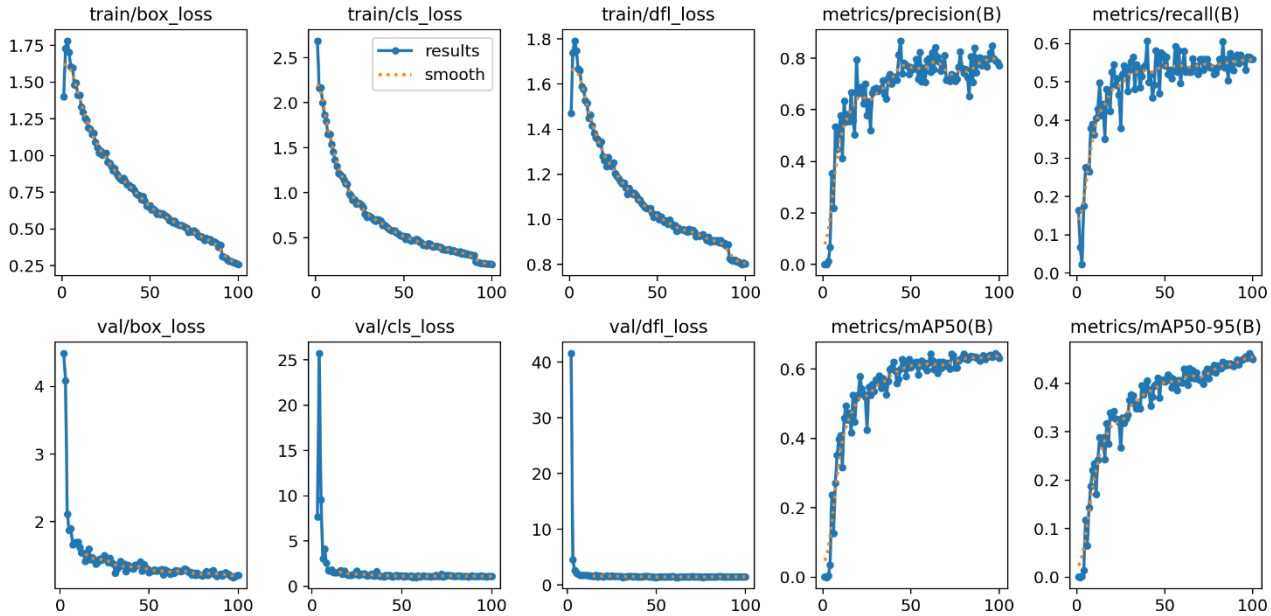
Annex 12: D3-BLR Detailed Results

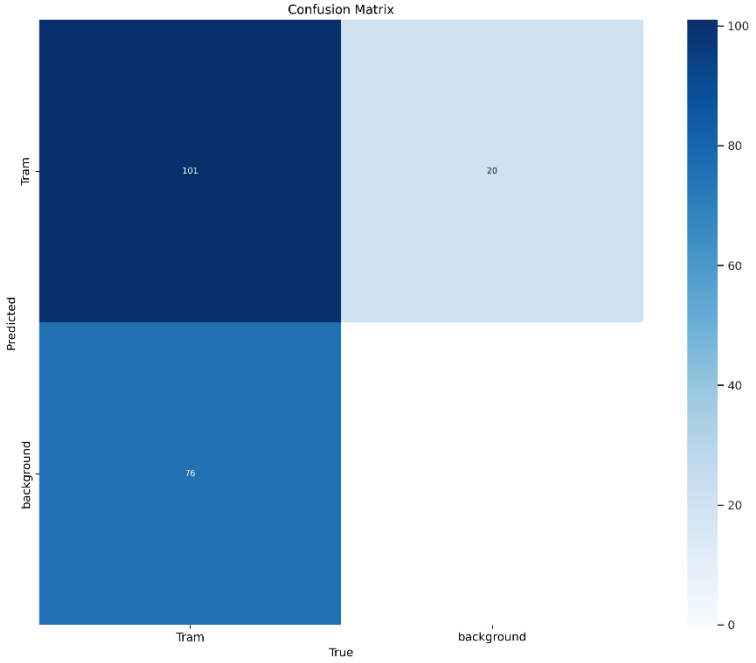


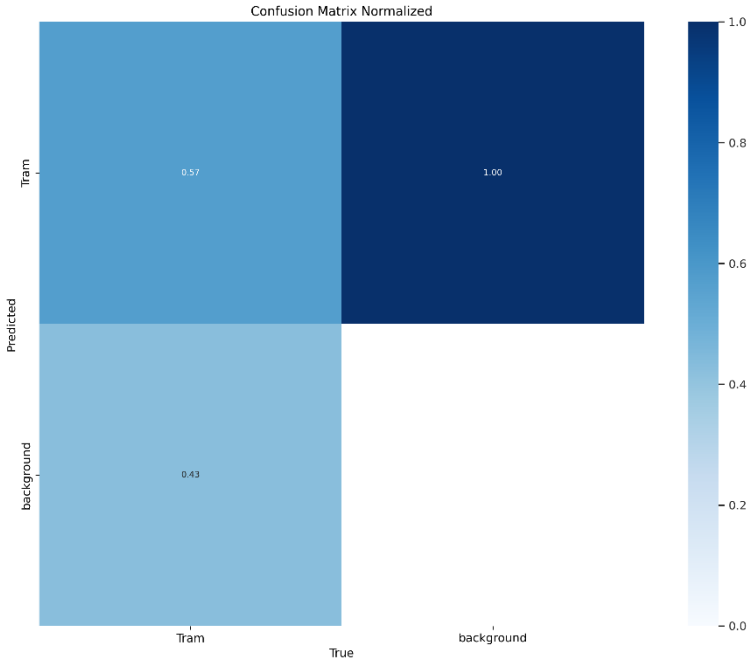




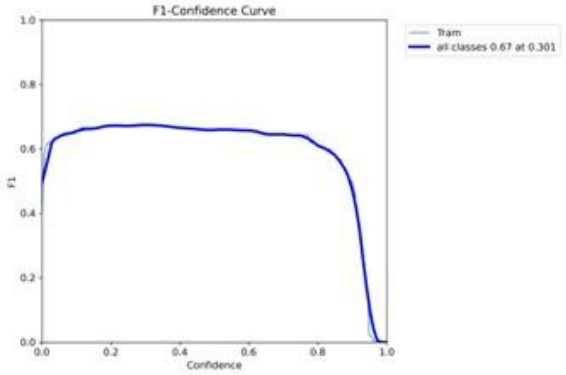


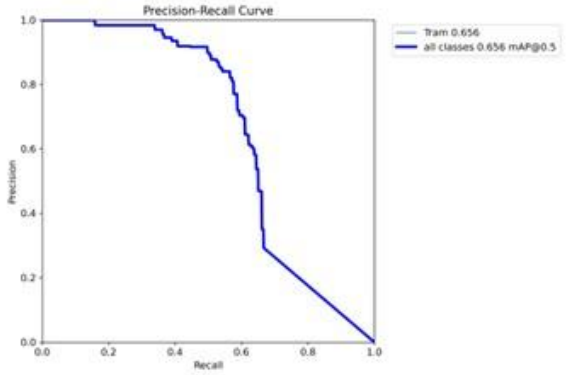


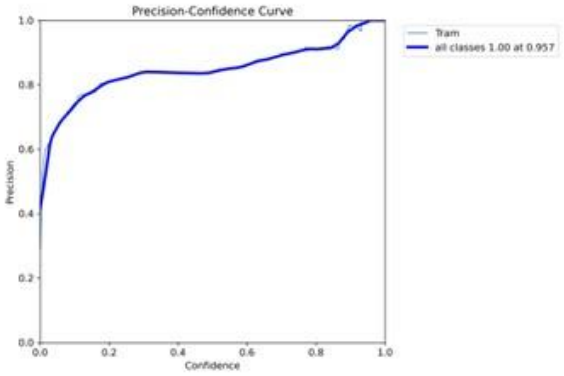

Confusion Matrix
Tram
background
Predicted
101
20
76
Tram
background
True
100
80
60
40
20
0

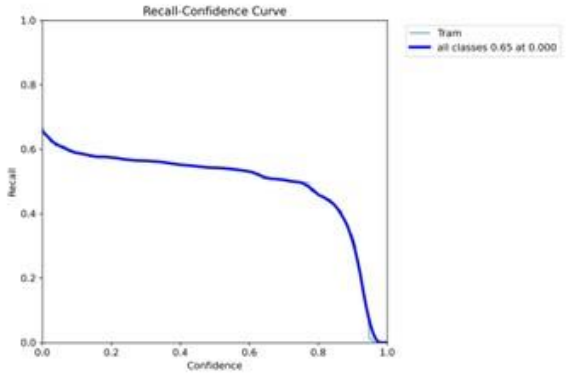

Confusion Matrix Normalized
Tram
background
Predicted
0.57
1.00
0.43
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

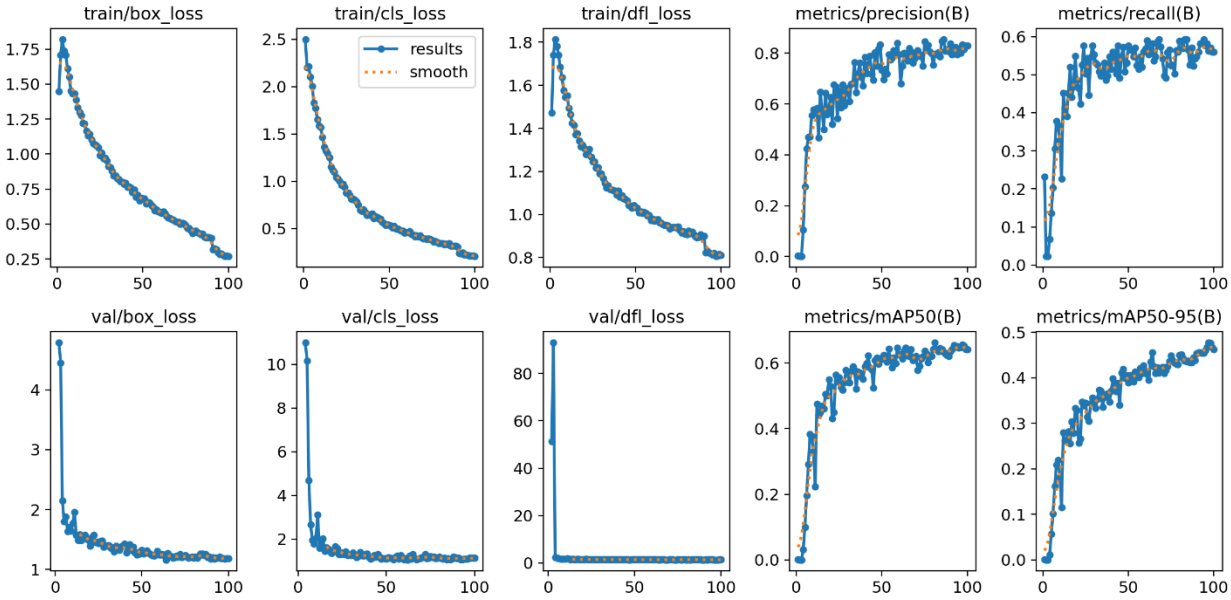
Annex 13: D4-SAT Detailed Results

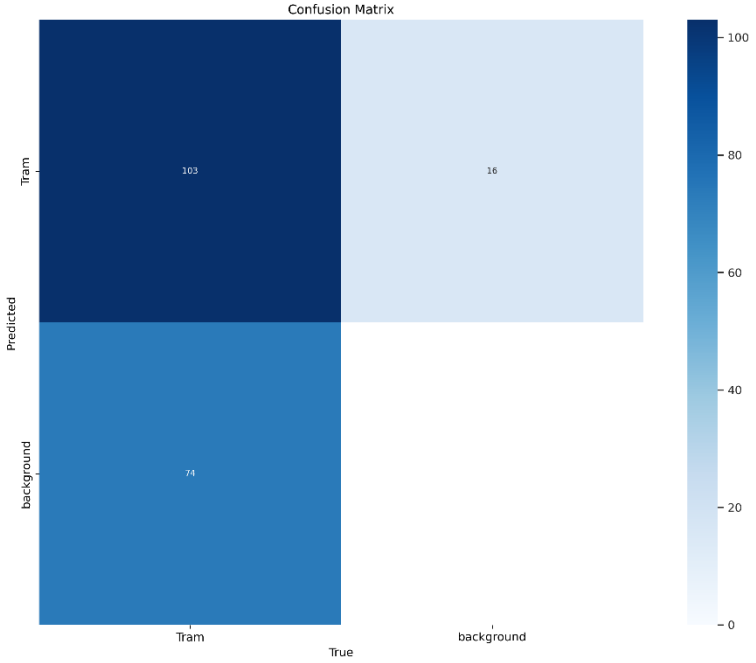


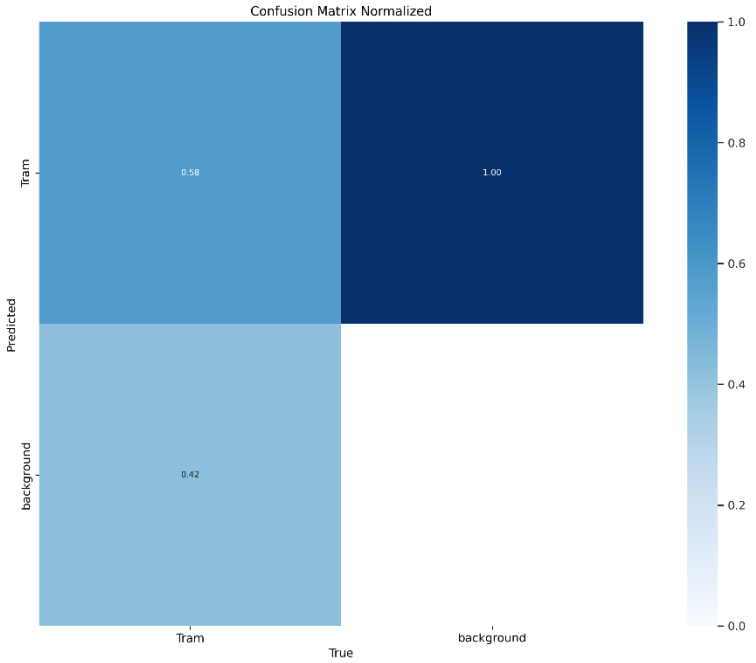




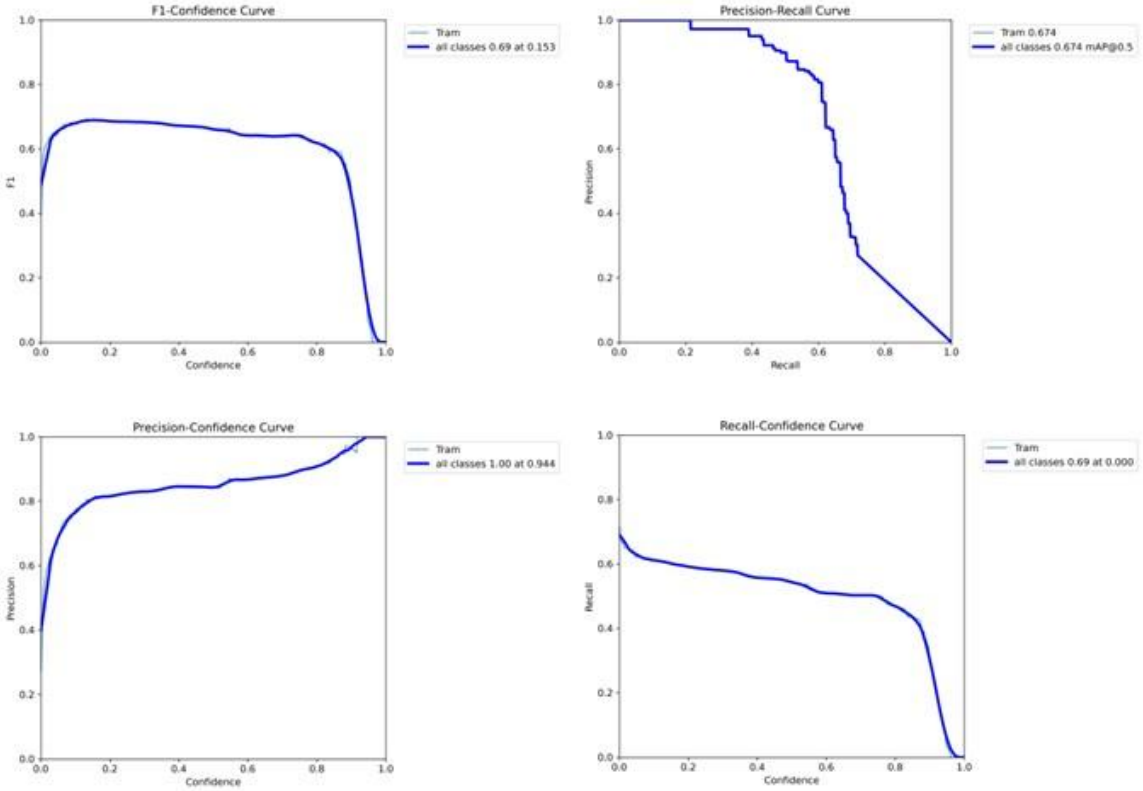


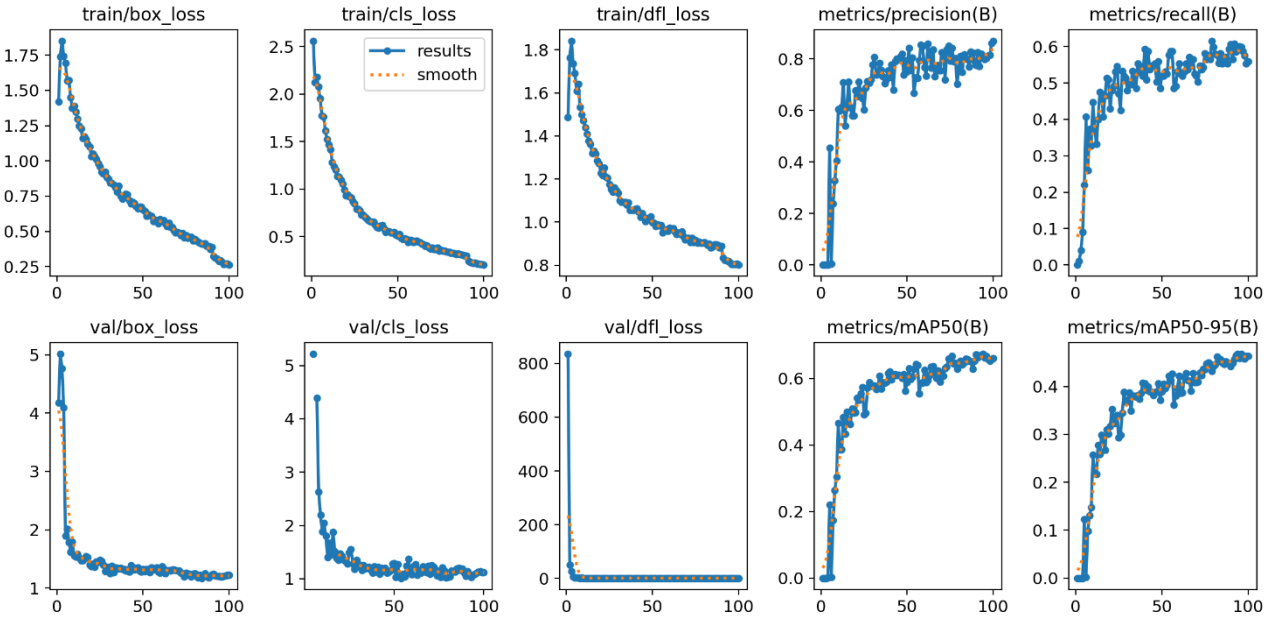


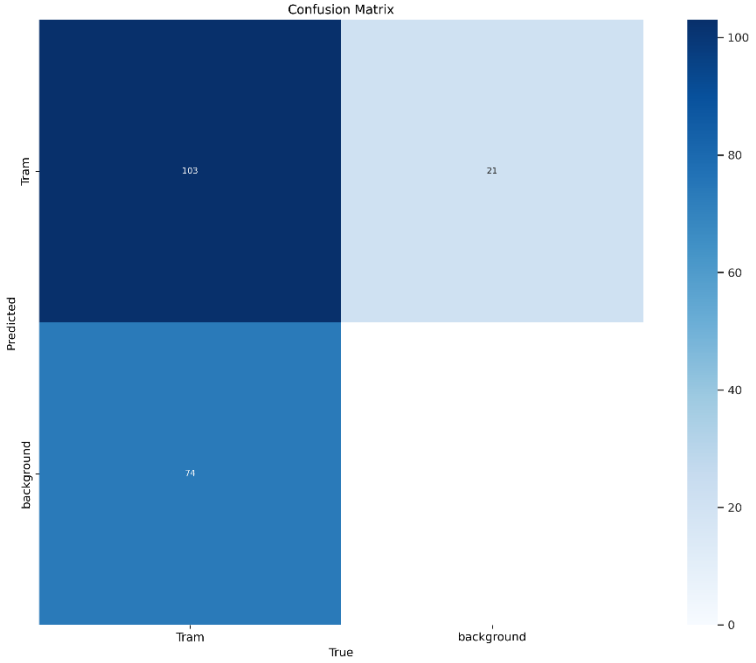

Confusion Matrix
Tram
background
Predicted
103
16
74
Tram
background
True
100
80
60
40
20
0

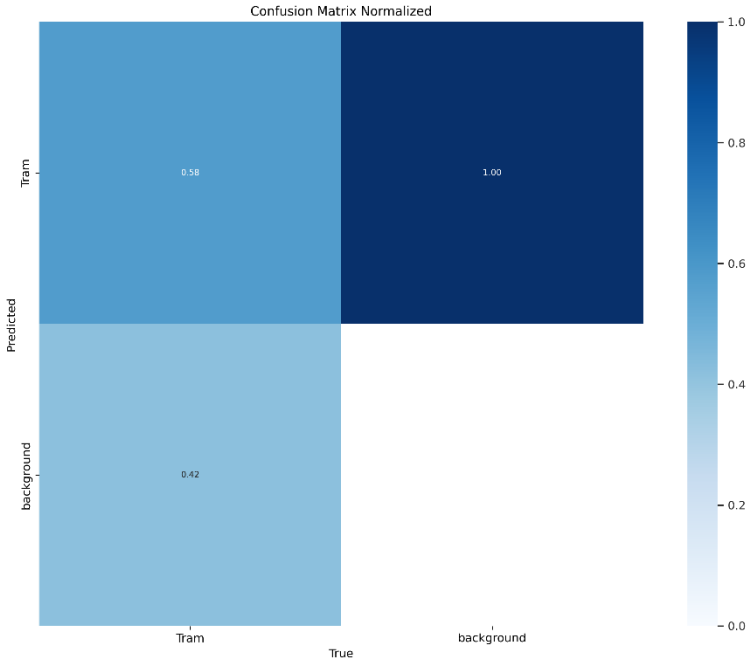

Confusion Matrix Normalized
Tram
background
Predicted
0.58
1.00
0.42
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

Annex 14: D5-NS Detailed Results

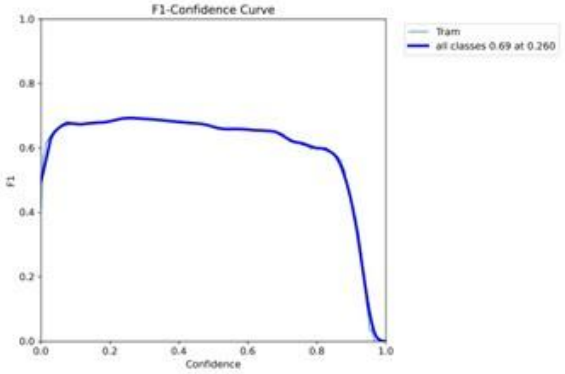


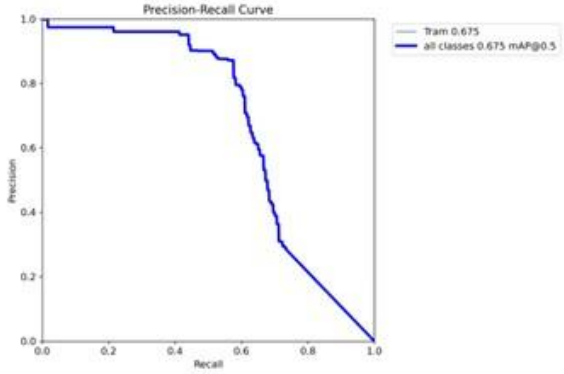


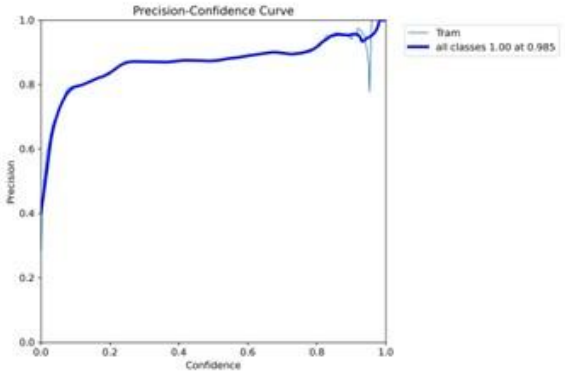

Confusion Matrix
Tram
background
Predicted
103
21
74
Tram
background
True
100
80
60
40
20
0

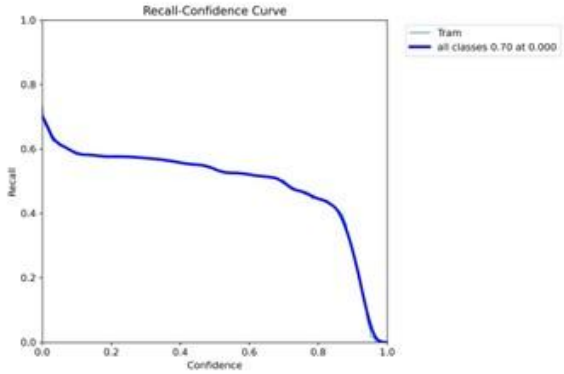

Confusion Matrix Normalized
Tram
background
Predicted
0.58
1.00
0.42
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

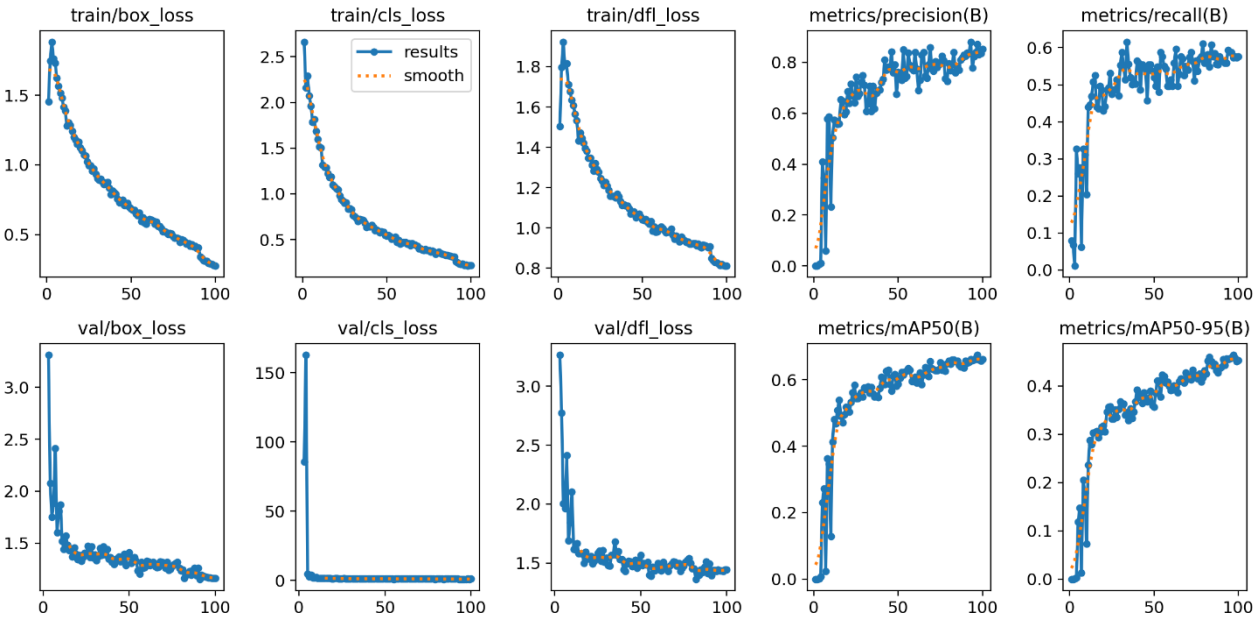
Annex 15: D6-Allx3 Detailed Results

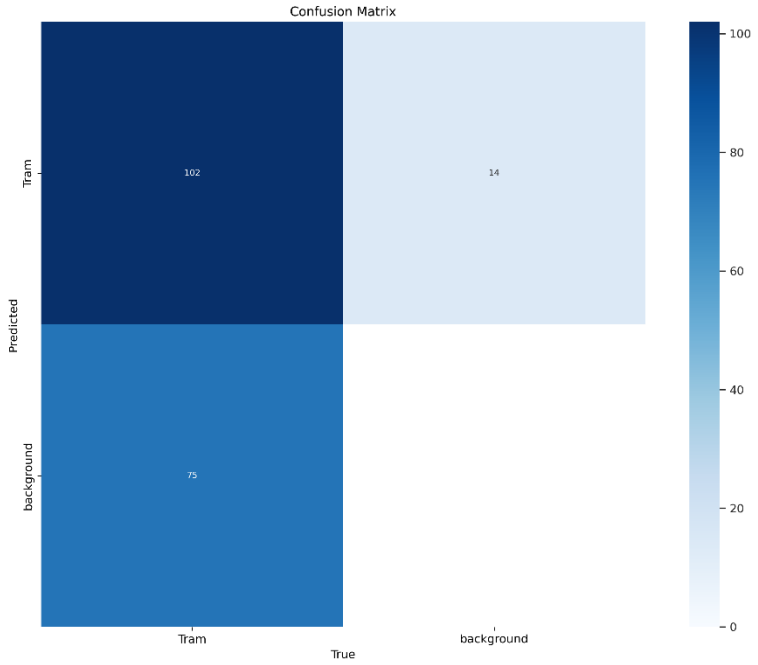


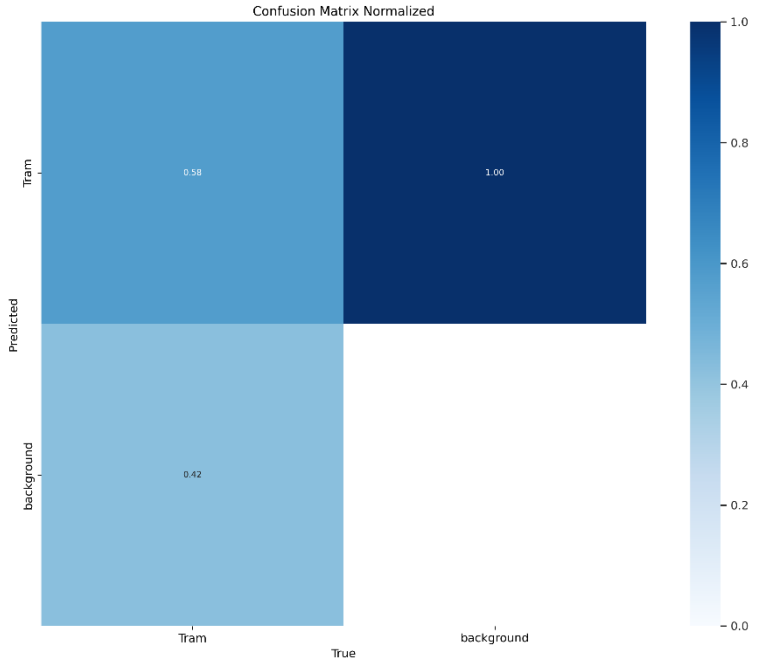




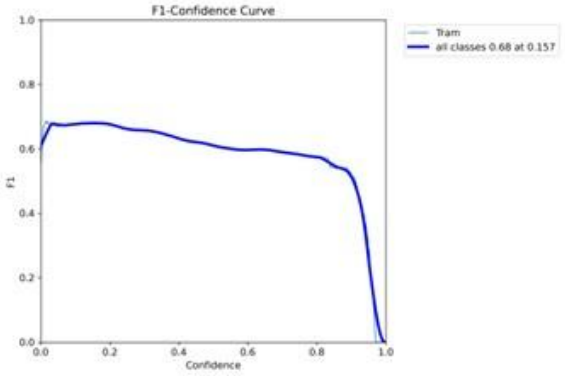


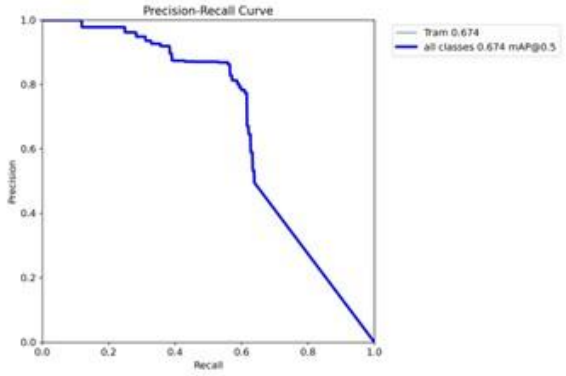


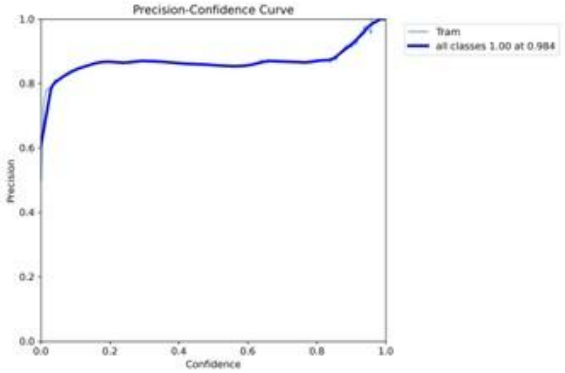

Confusion Matrix
Tram
background
Predicted
102
14
75
Tram
background
True
100
80
60
40
20
0

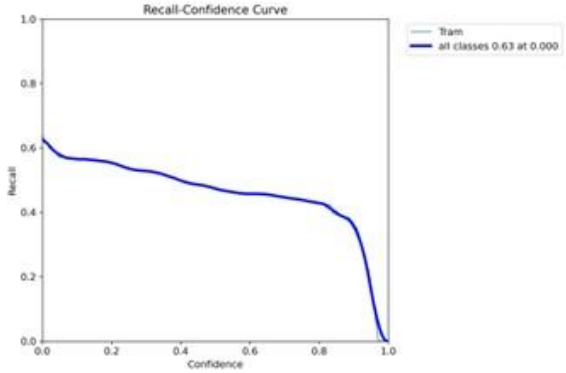

Confusion Matrix Normalized
Tram
background
Predicted
0.58
1.00
0.42
Tram
background
True
1.0
0.8
0.6
0.4
0.2
0.0

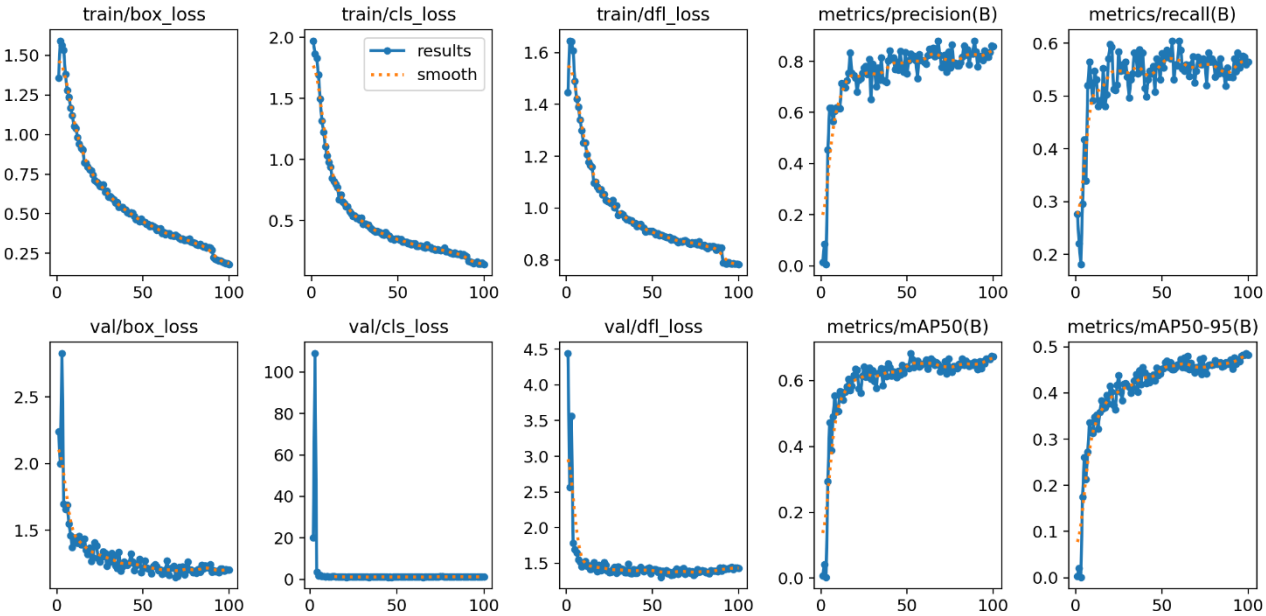
Annex 16: D7-Allx7 Detailed Results

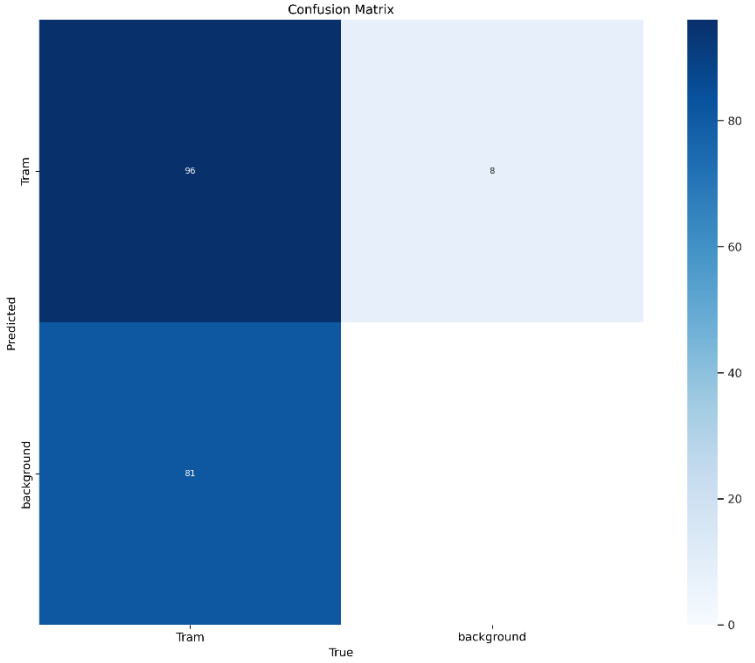


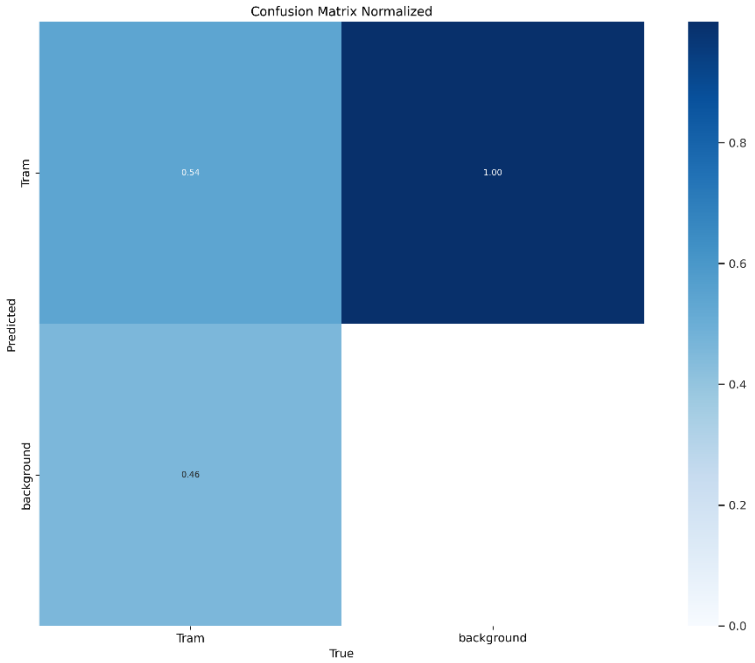









Confusion Matrix
Tram
background
Predicted
96
8
81
Tram
background
True
80
60
40
20
0

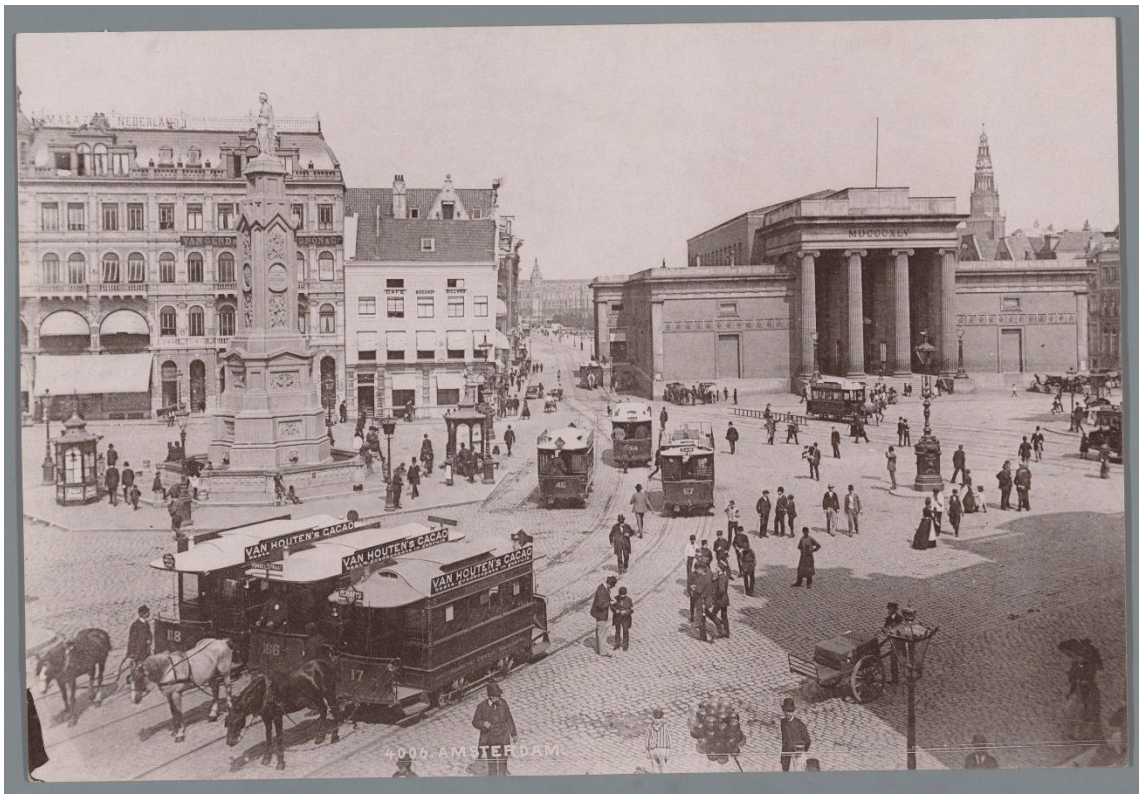

Confusion Matrix Normalized
Tram
background
Predicted
0.54
1.00
0.46
Tram
background
True
0.8
0.6
0.4
0.2
0.0

Annex 17: Original photograph Figure 14

Annex 18: Original photograph Figure 15

Annex 19: Original photograph Figure 16

Annex 20: Original photograph Figure 17

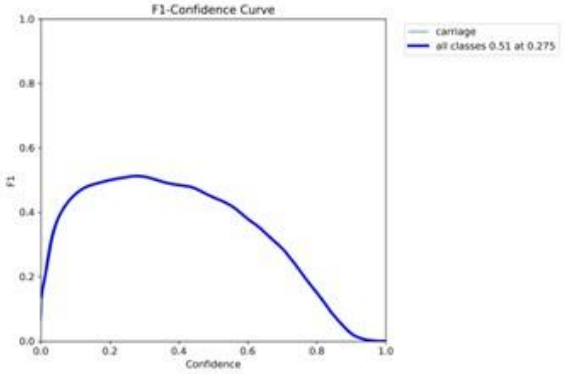


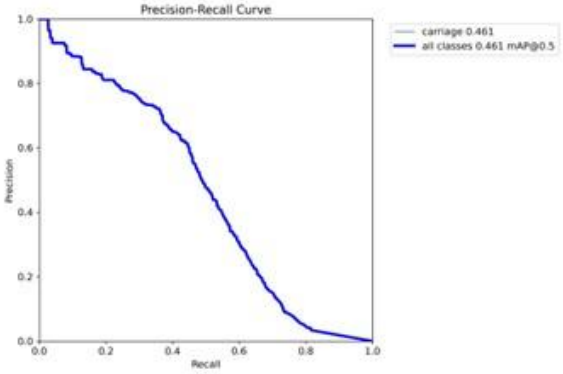
Annex 21: Original photograph Figure 18

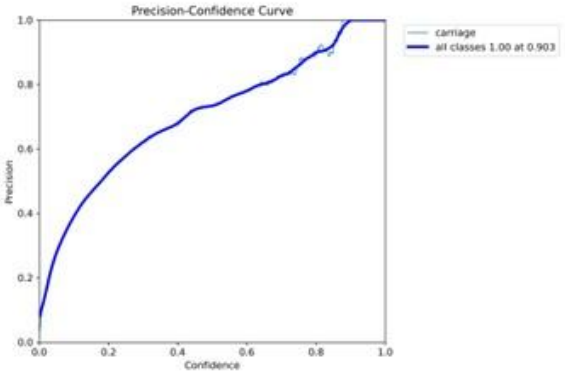
Annex 22: Original photograph Figure 19

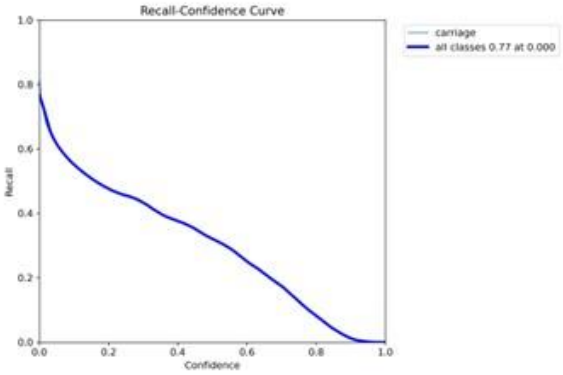
Annex 23: Original photograph Figure 20

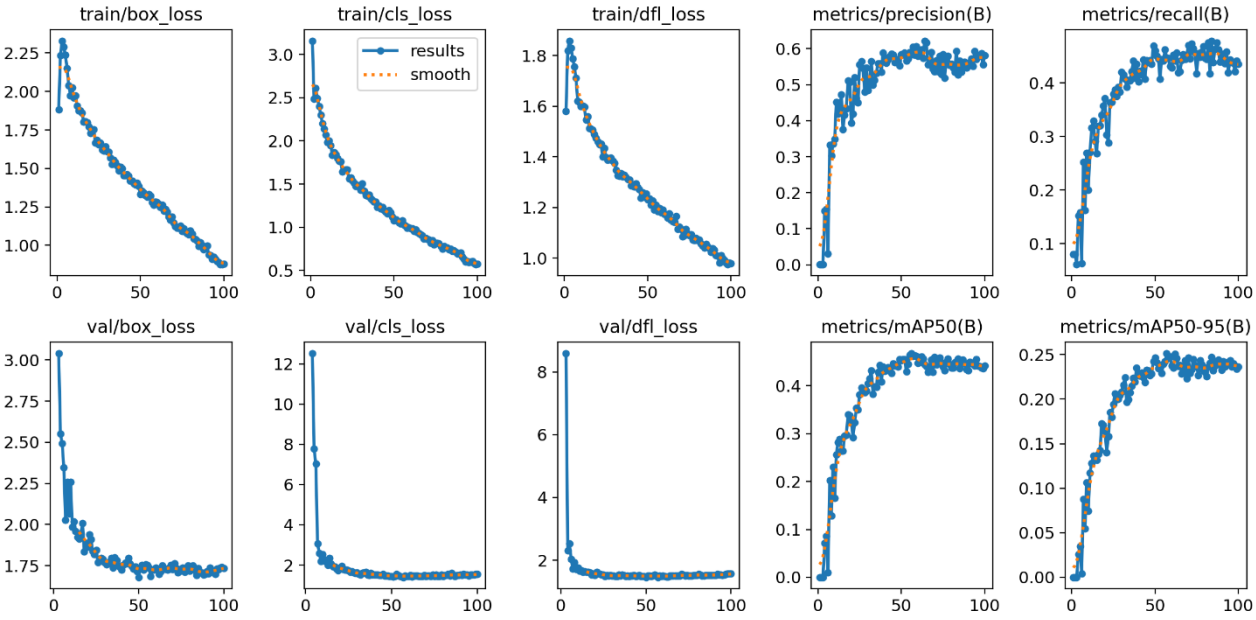
Annex 24: D0-Base Detailed Results

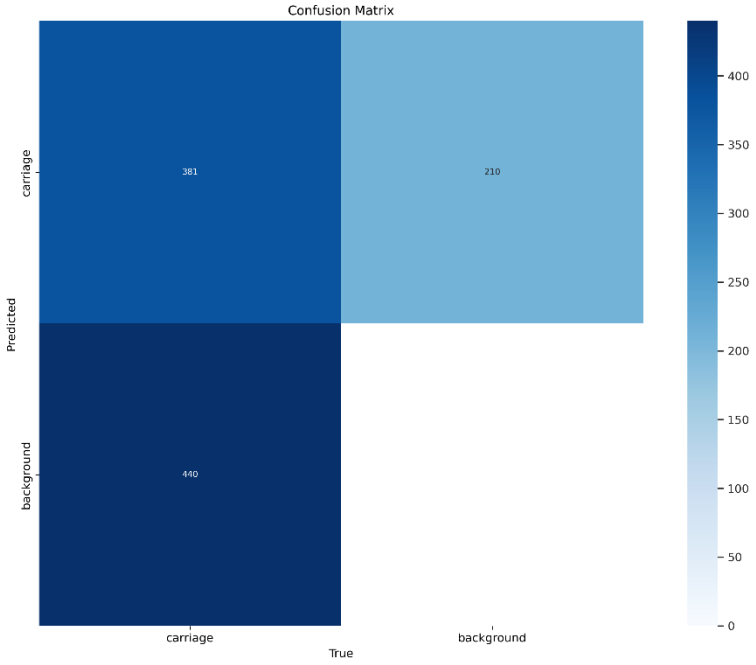


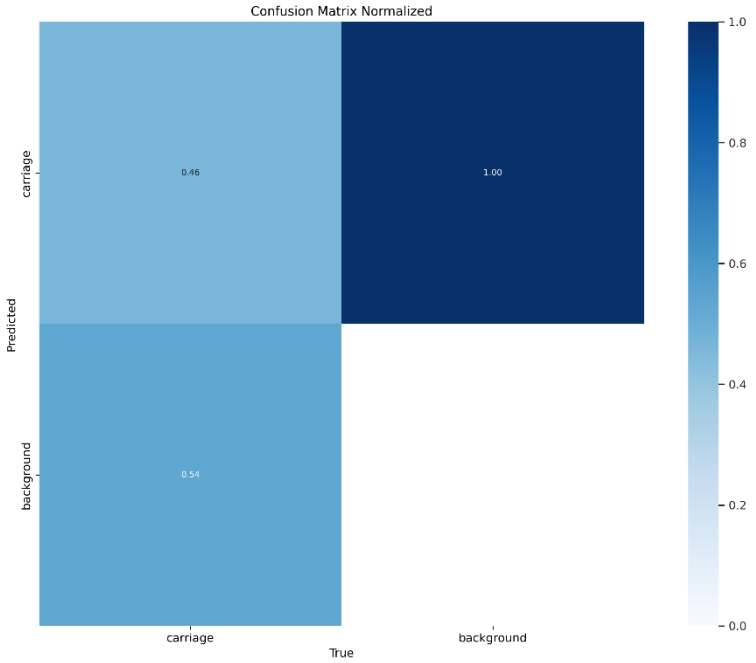




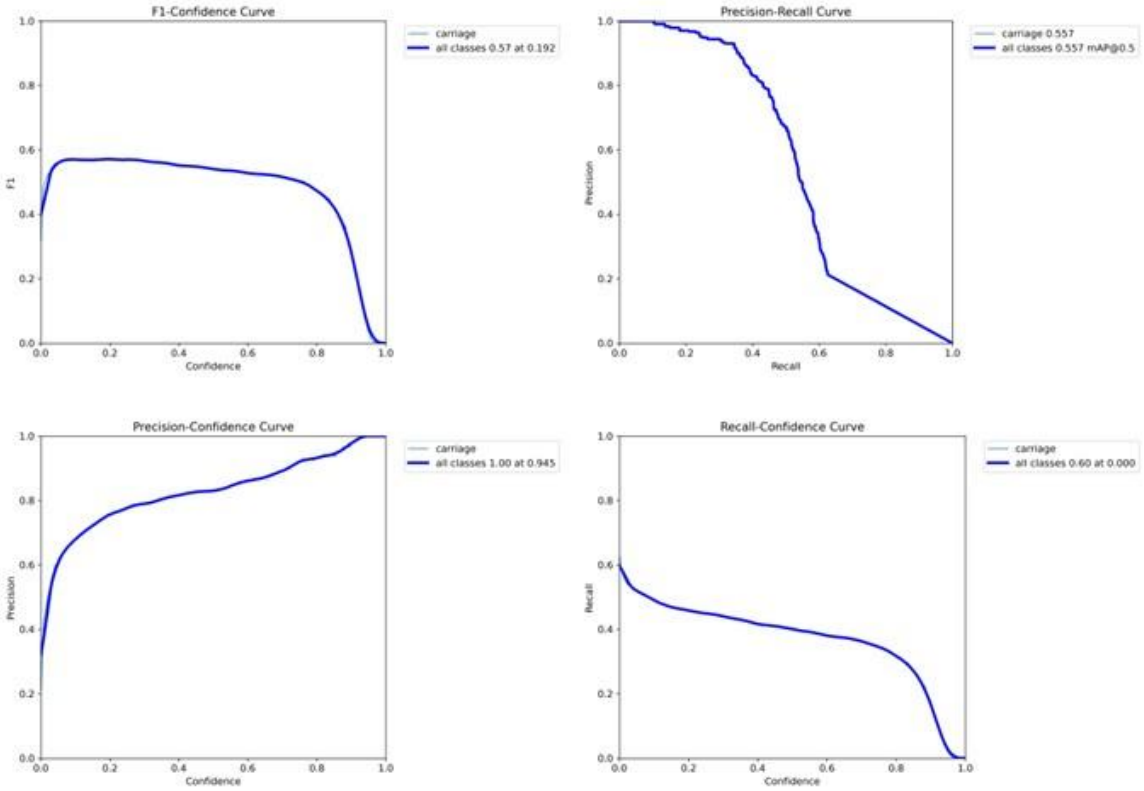


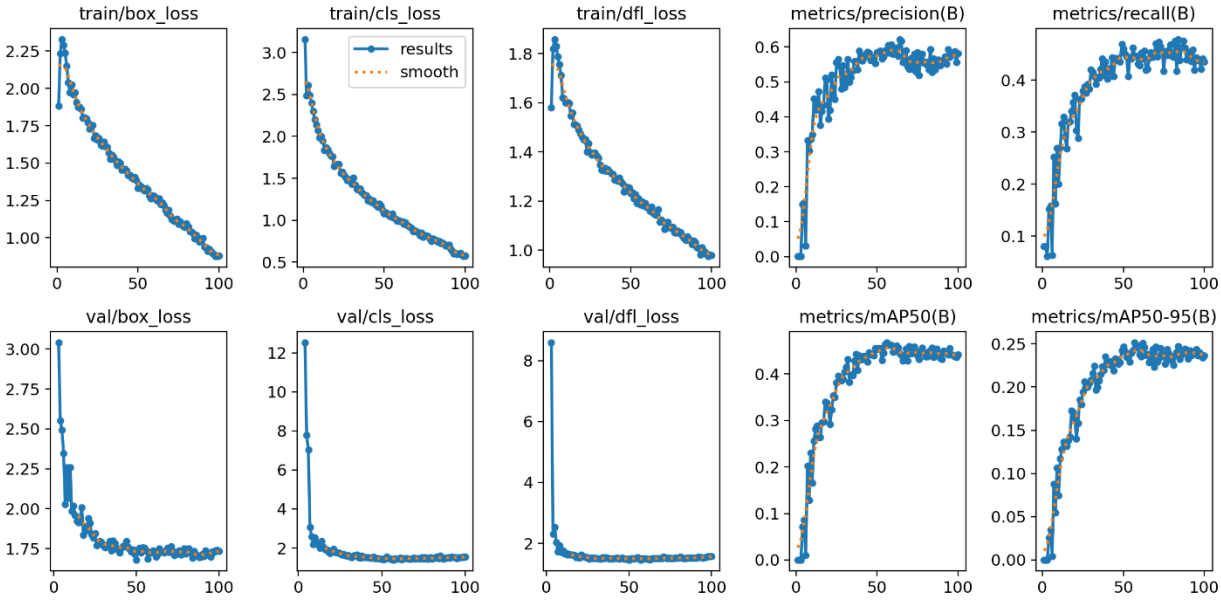


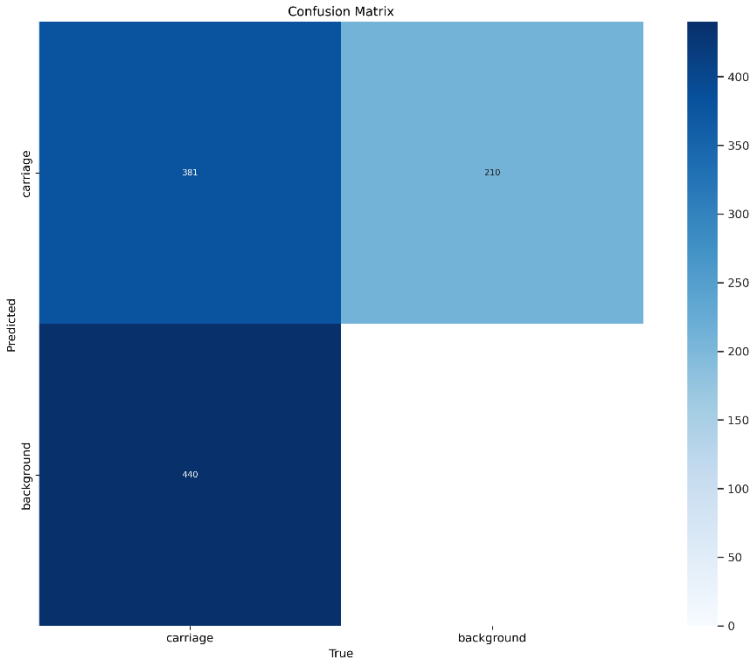

Confusion Matrix
381
210
440
carriage
background
Predicted
carriage
background
True
400
350
300
250
200
150
100
50
0

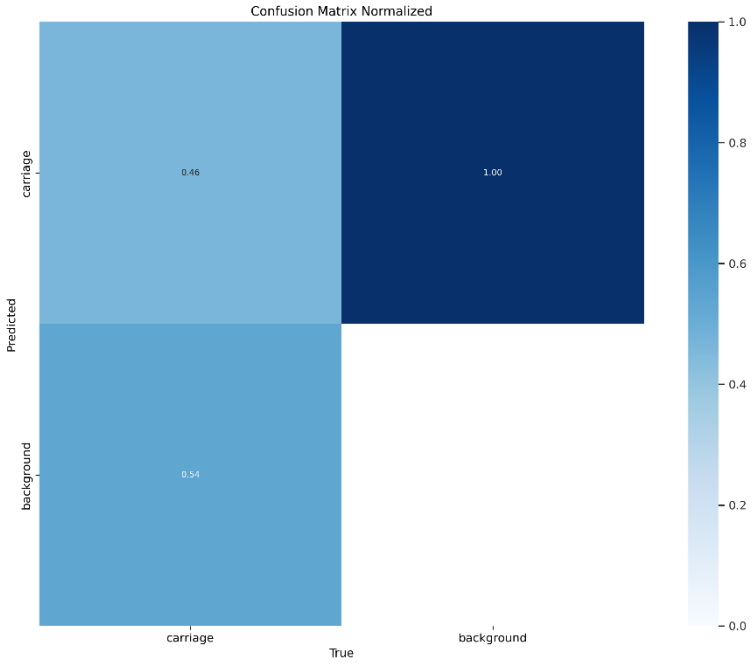

Confusion Matrix Normalized
0.46
1.00
0.54
carriage
background
Predicted
carriage
background
True
1.0
0.8
0.6
0.4
0.2
0.0

Annex 25: D1-Allx7 Detailed Results





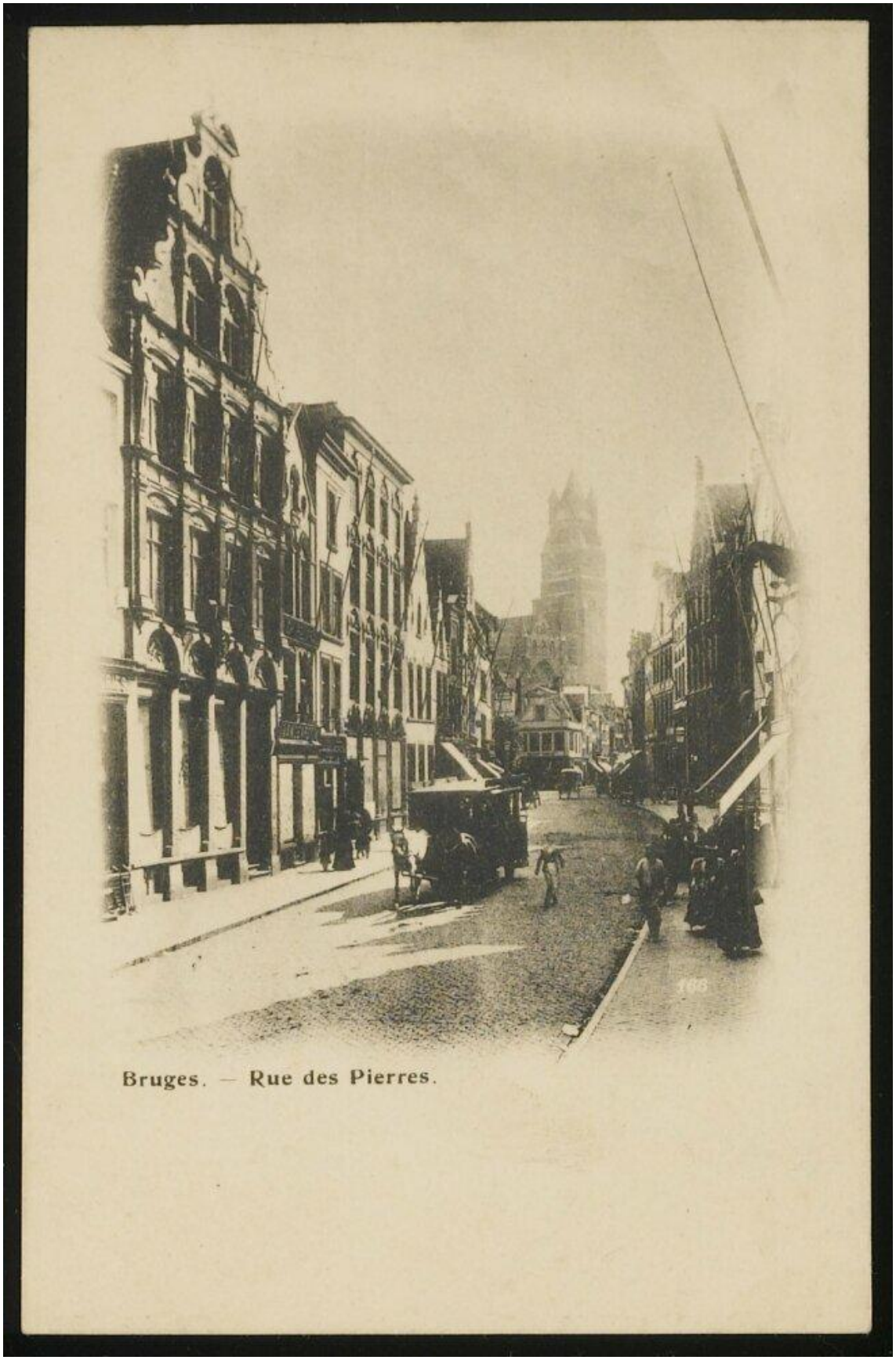

Confusion Matrix
carriage
background
Predicted
381
210
440
carriage
background
True
400
350
300
250
200
150
100
50
0


Confusion Matrix Normalized
carriage
background
Predicted
0.46
1.00
0.54
carriage
background
True
1.0
0.8
0.6
0.4
0.2
0.0

Annex 26: Original photograph Figure 22

Annex 27: Original photograph Figure 23

Annex 28: Original photograph Figure 24

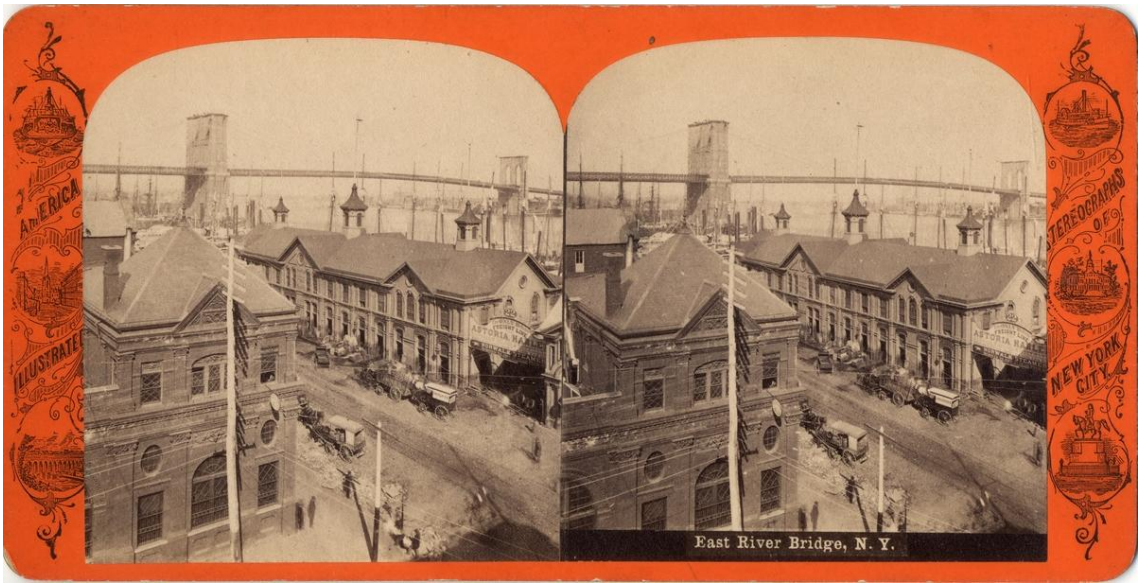


Annex 29: Original photograph Figure 25

Annex 30: Original photograph Figure 26

Annex 31: Original photograph Figure 27

Annex 32: Original photograph Figure 28



Annex 33: Original photograph Figure 29

Annex 34: Original photograph Figure 30

Annex 35: Original photograph Figure 31

Annex 36: Original photograph Figure 32

Annex 37: Discriminator, Generator, Cycle Consistency and Identity Losses