UNIVERSITY OF
LATVIA

Summary of
Doctoral Thesis

Maksims Ivanovs

DEEP LEARNING FOR APPLIED COMPUTER VISION: SOLVING IMAGE UNDERSTANDING TASKS WITH CONVOLUTIONAL NEURAL NETWORKS

Riga 2025

UNIVERSITY OF
LATVIA

FACULTY OF SCIENCE
AND TECHNOLOGY

Maksims Ivanovs

DEEP LEARNING FOR APPLIED COMPUTER VISION: SOLVING IMAGE UNDERSTANDING TASKS WITH CONVOLUTIONAL NEURAL NETWORKS

SUMMARY OF DOCTORAL THESIS

Submitted for the Doctoral degree (Ph. D.)
in Computer Science
Subfield of Artificial Intelligence

Riga 2025

The doctoral thesis was carried out at the Institute of Electronics and Computer Science (EDI) and the Department of Computing of the Faculty of Science and Technology, University of Latvia, from 2019 to 2024.

SAM 8.2.2. 'Akadēmiskā personāla atjaunotne un kompetenču pilnveide Latvijas Universitāte', projekta Nr. 8.2.2.0/18/A/010.

The thesis contains the introduction, five chapters, the conclusion, and the list of references.

Form of the thesis: dissertation in Computer Science, in the subfield of Artificial Intelligence.

Scientific Advisor: Senior Researcher, Dr. sc. ing. **Roberts Kadiķis**.

Reviewers:

1. Professor, Dr. sc. comp. **Jānis Zuters**, University of Latvia.
2. Associate Professor, Dr. sc. ing. **Andrejs Romānovs**, Riga Technical University.
3. Lecturer, PhD **Mahmoud Elbattah**, College of Arts, Technology and Environment, University of the West of England, Bristol, United Kingdom and Laboratoire MIS, Université de Picardie Jules Verne, Amiens, France.

The thesis will be defended at the public session of the Doctoral Committee of Computer Science and Informatics, University of Latvia, on 16 May 2025.

The thesis is available at the Library of the University of Latvia, Kalpaka blvd. 4.

Chairman of the Doctoral Committee _____ / **Guntis Bārzdiņš** /

Secretary of the Doctoral Committee _____ / **Sintija Siliņa** /

ISBN 978-9934-36-374-0
ISBN 978-9934-36-375-7 (PDF)

Abstract

This PhD thesis focuses on the applications of deep learning methods to solving key image understanding tasks: image classification, object detection, and semantic segmentation.

In the literature review in Chapter 1, I provide the background for the experimental work presented in the rest of the thesis, establish that the most promising type of deep neural architecture for image understanding tasks is convolutional neural networks (CNNs), discuss challenges related to the availability of data for training CNNs, and outline how transfer learning and synthetic data can be leveraged to mitigate these challenges.

The practical part of the thesis focuses on the applications of CNNs to real-world image understanding problems. In Chapter 2, I detail the research on the use of CNNs for recognising hand-washing movements, aiming to design a system to monitor hand hygiene. In Chapter 3, I describe the use of CNNs for semantic segmentation of street views, which is an essential task for navigation systems of self-driving cars. In Chapter 4, I describe the use of CNN-based object detectors for identifying plastic bottles that can be picked up by a robotic arm. Finally, in Chapter 5, I address the application of CNNs to the task of classifying microscopy images with the goal of automating the monitoring of the growth of organs-on-a-chip.

The goal of the thesis – to provide efficient solutions for applied image understanding tasks – was achieved for all tasks except the classification experiments on the PSKUS dataset, the most complex and noisy dataset of hand-washing videos. Additionally, the findings that I report in this thesis contributed to the better understanding of methodological challenges in deep learning such as approaches to augmenting real-world datasets with synthetic data. The results reported in the thesis have been published in six scientific articles indexed in Elsevier Scopus and/or Web of Science databases and presented at four conferences. The approbation of the results was conducted in seven research projects at the Institute of Electronics and Computer Science (EDI), where this thesis was carried out. The results support the four thesis statements that I propose for the defence.

Contents

Introduction

The field of computer vision, which began to develop in the late 1960s, is a vast and one of the most active research areas in computer science today, as there are many possible applications for automated systems capable of understanding visual scenes. In this thesis, I am concerned with three types of image understanding problems: image classification, image segmentation, and object detection. My work on these problems belongs to the domain of applied computer vision, meaning the application of computer vision methods to practical tasks in science and industry. In particular, I use computer vision methods to solve the following real-world image understanding tasks:

- to classify hand-washing movements in clinical settings for monitoring hand hygiene (Chapter 2);
- to perform semantic segmentation of street views for improving the navigation systems of self-driving cars (Chapter 3);
- to detect graspable bottles in a pile for the bin-picking task carried out by a robotic arm (Chapter 4);
- to classify microscopy images for automated monitoring of growing organs-on-a-chip (OOC) (Chapter 5).

In all studies presented in this thesis, I address image understanding challenges by means of deep learning (DL), i.e., using deep neural networks (DNN), which are currently regarded as the state-of-the-art for most computer vision tasks. Given that DNNs have also been successfully applied in many other artificial intelligence (AI) domains, addressing general methodological challenges in DL in this thesis – such as the need for large datasets for training models, quality control issues in large-scale datasets, and the use of synthetic data for augmenting real-world datasets – can contribute to the advancements in these domains as well.

The goal of this thesis is to provide efficient solutions for applied image understanding tasks. **The central premise** of the thesis is that convolutional neural networks (CNNs), a type of DNNs particularly suitable for processing perceptual data, can successfully solve the image understanding tasks considered in this work. **The research objectives** and **hypotheses** depend on the specific task and therefore are defined in the chapters of the thesis reporting respective studies. **The research methods** that I use in this thesis are those commonly employed in AI and computer vision research: exploration and analysis of relevant literature, data cleaning and preprocessing, synthetic data generation, design and implementation of experiments involving DNNs, and analysis and validation of the results

of experiments.

As a result of the work presented in this thesis, I propose the following **thesis statements** for defence:

- **Thesis statement one:** In applications of CNNs to real-world image understanding tasks, data availability and quality present greater challenges than model selection and customisation.
- **Thesis statement two:** CNNs that perform well when trained and evaluated on datasets acquired in laboratory conditions may struggle to achieve similar success when trained and evaluated on more complex real-world data.
- **Thesis statement three:** While state-of-the-art CNN-based image classifiers and object detectors with a larger number of parameters typically demonstrate higher accuracy on benchmark datasets than their counterparts with a smaller number of parameters, this accuracy gap narrows or even vanishes when these models are trained and evaluated on smaller, more complex real-world datasets.
- **Thesis statement four:** While augmenting real-world datasets with photorealistic synthetic images is an efficient way to improve the accuracy of CNNs trained on such data, increasing the amount of synthetic data does not directly correlate with improved accuracy on image understanding tasks.

The above thesis statements are primarily grounded in the results found in the following chapters: thesis statement one – in Chapters 2, 3, 4, and 5; thesis statement two – in Chapter 2; thesis statement three – in Chapters 4 and 5; thesis statement four – in Chapters 3 and 5.

Research findings reported in this thesis have been published in the following scholarly articles indexed in Elsevier Scopus or/and Web of Science databases:

1. M. Lulla, A. Rutkovskis, A. Slavinska, A. Vilde, A. Gromova, **M. Ivanovs**, A. Skadins, R. Kadikis, and A. Elsts, "Hand-washing video dataset annotated according to the World Health Organization's hand-washing guidelines," *Data*, vol. 6, no. 4:38, 2021.
2. A. Elsts, **M. Ivanovs**, R. Kadikis, and O. Sabelnikovs, "CNN for hand washing movement classification: What matters more – the approach or the dataset?," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, IEEE, 2022.
3. **M. Ivanovs**, K. Ozols, A. Dobrajs, and R. Kadikis, "Improving semantic segmentation of urban scenes for self-driving cars with synthetic images," *Sensors*, vol. 22, no. 6:2252, 2022.

4. D. Duplevska, **M. Ivanovs**, J. Arents, and R. Kadikis, "Sim2Real image translation to improve a synthetic dataset for a bin picking task," in *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1–7, IEEE, 2022.
5. **M. Ivanovs**, L. Leja, K. Zviedris, R. Rimsa, K. Narbute, V. Movcana, F. Rumnieks, A. Strods, K. Gillois, G. Mozolevskis, A. Abols, and R. Kadikis, "Synthetic image generation with a fine-tuned latent diffusion model for organ on chip cell image classification," in *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp. 1–6, IEEE, 2023.
6. V. Movčana, A. Strods, K. Narbute, F. Rūmnieks, R. Rimša, G. Mozoļevskis, **M. Ivanovs**, R. Kadiķis, K. Zviedris, L. Leja, A. Zujeva, T. Laimiņa, and A. Abols, "Organ-On-A-Chip (OOC) Image Dataset for Machine Learning and Tissue Model Evaluation," *Data*, vol. 9, no. 2:28, 2024.

Research findings included in this thesis have also been reported in the following scholarly publications that are not indexed in Elsevier Scopus or Web of Science databases:

1. **M. Ivanovs**, R. Kadikis, M. Lulla, A. Rutkovskis, and A. Elsts, "Automated quality assessment of hand washing using deep learning," *arXiv:2011.11383*, 2020.
2. O. Zemlanuhina, M. Lulla, A. Rutkovskis, A. Slavinska, A. Vilde, A. Melbarde-Kelmere, A. Elsts, **M. Ivanov**, and O. Sabelnikovs, "Influence of different types of real-time feedback on hand washing quality assessed with neural networks/simulated neural networks," in *SHS Web of Conferences*, vol. 131, p. 02008, EDP Sciences, 2022.

I have presented research findings reported in this thesis at the following conferences:

1. IEEE International Conference on Image Processing Theory, Tools & Applications – IPTA 2022, Salzburg, Austria, 2022.
   Presentation *CNN for Hand Washing Movement Classification: What Matters More — the Approach or the Dataset?*
2. 27$^{\text{th}}$ IEEE International Conference on Emerging Technologies and Factory Automation – EFTA 2022, Stuttgart, Germany, 2022.
   Presentation *Sim2Real Image Translation to Improve a Synthetic Dataset for a Bin Picking Task.*
3. SEMICON Europa 2022 Conference, Munich, Germany, 2022.
   Presentation *Synthetic Data for Robotics: Opportunities and Challenges.*

4. 26th IEEE Signal Processing: Algorithms, Architectures, Arrangements and Applications Conference – SPA 2023, Poznan, Poland, 2023.
   Presentation *Synthetic Image Generation With a Fine-Tuned Latent Diffusion Model for Organ on Chip Cell Image Classification*.

The research for this thesis was conducted and approbated at the Institute of Electronics and Computer Science (EDI – *Elektronikas un datorzinātņu institūts*). The research was part of several scientific projects at EDI and was financially supported by their funding. The following is the list of these projects:

1. *Programmable Systems for Intelligence in Automobiles – PRYSTINE* (Horizon 2020 ECSEL Joint Undertaking funding under grant agreement 783190).
2. *Efficient module for automatic detection of people and vehicles using video surveillance cameras – VAPI* (ERDF project No. 1.2.1.1/18/A/006 research No. 1.5).
3. *Automated hand washing quality control and hand washing quality evaluation system with real-time feedback – Handwash* (project No. lzp-2020/2-0309).
4. *Integration of reliable technologies for protection against Covid-19 in healthcare and high risk areas – COV-CLEAN* (project No. VPP-COVID-2020/1-004).
5. *Intelligent Motion Control under Industry 4.E – IMOCO4.E* (Horizon 2020 ECSEL Joint Undertaking funding under grant agreement 101007311).
6. *AI-Improved Organ on Chip Cultivation for Personalised Medicine – AImOOC*
   (contract with Central Finance and Contracting Agency of Republic of Latvia no. 1.1.1.1/21/A/079; the project was cofinanced by REACT-EU funding for mitigating the consequences of the pandemic crisis).
7. *Holographic microscopy- and artificial intelligence-based digital pathology for the next generation of cytology in veterinary medicine – VetCyto* (project No. lzp-2023/1-0220).

The thesis consists of the introduction, five chapters, the conclusion, and the list of references. The length of the thesis is 146 pages.

1. Background

In this thesis, I adopt the definition of **computer vision** given in [1] as 'the host of techniques to acquire, process, analyze, and understand complex higher-dimensional [visual] data from our environment'. I differentiate it from digital image processing, following the approach in [2] and maintaining that digital image processing refers to low-level operations on images such as image enhancement or colour processing, whereas the broader term 'computer vision' also includes higher-level operations with visual data.

As computer vision spans multiple disciplines and is concerned with a wide range of research problems, I introduce a narrower term to refer to the research problems that my work is primarily concerned with. For that purpose, I use the term **image understanding** as defined in [3], that is, as a suite of methods and techniques that 'attempt to interpret the meaning of image at a high level to provide semantic information closely related to human thinking, and help further to make decisions and to guide the actions according to the understanding of scenes'. My research that forms the foundation of the present thesis focuses on three key image understanding tasks: **image classification**, **object detection**, and **image segmentation**.

Image classification is essentially a labelling procedure [4]: a classifier labels a given image I as belonging to a single class C_i, which is an element of the fixed set of considered classes $C = \{C_1, C_2, C_3, ..., C_N\}$.

Object detection is a more complex task than image classification, because an object detector needs to both assign each object in the image a label $C_i \in C$ and identify the positions of the objects. In other words, while an image classifier aims to answer the question *What object is there?*, an object detector needs to answer the question *What objects are where?* [5].

Image segmentation aims at partitioning an input image into multiple segments (i.e., continuous groups of pixels) in a meaningful way. One of the main types of image segmentation, **semantic segmentation**, has the goal of labelling each pixel of the resulting segments with a single class label $C_i \in C$. Therefore, semantic segmentation is essentially a multiclass classification at the pixel level; as it is performed for each pixel of an input image, it is generally considered a more challenging task than image classification [6].

1.1 Methods for solving image understanding tasks

Methods for solving image understanding tasks can broadly be divided into two principal categories: classical methods, and DL-based methods.

The defining characteristic of **classical methods** for image understanding is that they are fully or partially based on explicitly programmed algorithms and rely on hand-crafted features and rules designed by human experts in specific domains. While these methods are still sometimes used nowadays, DNNs are usually preferred over classical approaches [7], since they are more accurate and robust.

The rise in popularity of **DL-based methods** began in 2012, when the CNN AlexNet [8] won the ILSVRC image classification competition [9]. In just a few years, image classification became dominated by DL-based methods, and a similar shift occurred in object detection and semantic segmentation. As a result, DL is currently a vast and very rapidly evolving field of research in both academic and industrial domains, with a broad range of architectures available. **CNNs**, the most popular architecture for image understanding tasks, have revolutionised the use of artificial neural networks (ANNs) in computer vision [10], and are also the primary tool used in the research reported in this thesis. CNNs address the challenges of utilising spatial patterns in images more efficiently than simpler DNNs such as fully connected feedforward models; furthermore, the use of convolutions makes network connectivity more sparse, thereby improving efficiency.

The range of available CNN architectures is very broad, from the early pioneering LeNet [11] to the most recent state-of-the-art models. In the research presented in this thesis, I used MobileNetV2 [12], MobileNetV3Large [13], and EfficientNet-B7 [14] for image classification, Xception [15] and MobileNetV2 extended with a DeepLabv3 [16] segmentation head for semantic segmentation, and YOLOv5 [17] for object detection.

1.2 Datasets for image understanding tasks

The availability of large datasets for training, evaluating, and benchmarking models has been a major factor enabling the advancement of DL in general [18, 19] and in image understanding tasks in particular. Open access to these datasets has sparked competition to develop better models, democratised deep learning research and enabled researchers to work on image understanding tasks even without the resources to acquire and label data on their own. However, real-world applications often require specialised datasets, which leads to several data-related challenges when applying DL to real-world problems, such as collecting, curating, and labelling a sufficient amount of data and dealing with concept drift, that is,

the change between the class distribution at the time of training vs the current class distribution [20].

To mitigate the problems of data insufficiency and imbalance, one can use the a number of approaches, particularly:

- **transfer learning**, which aims to improve the performance of a model on a target domain by leveraging knowledge learned from another domain – the source domain [21];
- **fine-tuning**, which involves unfreezing previously frozen layers of the model and retraining them on the target dataset with a low learning rate;
- **data augmentation**, that is, extracting additional information for training a model from the original data by artificially inflating the training dataset by warping images or oversampling them [22];
- **using synthetic data**, that is, training a model on artificially generated data that are similar to the target domain.

2. Hand-Washing Movement Classification

Poor hand hygiene is a primary cause of the spread of multidrug-resistant bacteria and infections in clinical settings [23]. Compliance with hand hygiene recommendations reduces the prevalence of healthcare-associated infections [24], making adhering to protocols crucial. The World Health Organisation (WHO) provides widely adopted guidelines for hand washing [25], which outline six key movements. However, research [26, 27, 28] indicates that both the general public and medical professionals often neglect these steps, increasing the risk of spreading infections.

Monitoring hand washing is essential for improving compliance with the WHO guidelines and is traditionally done through direct observation [29, 30, 24]. Automated systems can offer continuous and more precise monitoring by capturing videos of hand-washing episodes, classifying movements, providing feedback to users, and storing data for further analysis. A promising approach is to use smartphones or edge devices as cost-efficient cores of such systems, making it easier to install monitoring systems in a hospital setting. In recent studies [31, 32, 33], promising results have been achieved with computer vision-based methods such as CNN-based classifiers, yet it is not clear whether the high accuracy achieved by such systems on lab-collected hand-washing datasets would translate well into successful performance in real-world environments. To address this uncertainty, the studies reported in this chapter involved experiments on more complex, real-world datasets. **The main hypothesis** was that lightweight CNNs, i.e., CNNs capable of running on mobile and edge devices, can successfully, i.e., with the accuracy above that of a putative 'naive' classifier, classify hand-washing movements in a real-world hospital setting.

2.1 Datasets

To address the lack of open access data for training ML-based hand-washing classifiers, two datasets – PSKUS and METC – were collected in a clinical setting and annotated.

2.1.1. PSKUS dataset

The PSKUS dataset was collected at Pauls Stradins Clinical University Hospital, using a custom IoT system. Cameras were installed above sinks in nine hospital locations, recording continuous hand-washing episodes.

In some locations, there was more than one camera installed to make recordings from different angles.

The labelling of the dataset was done by infectious disease experts and volunteers. The annotators first labelled each frame to indicate whether hand washing was present and subsequently to classify movements according to the WHO guidelines. The dataset contains 3 185 videos with 6 690 annotations, covering over 83 000 seconds of footage. Seven hand-washing movements were labelled per the WHO guidelines, along with the additional class 0 – 'other movement'.

Most videos were labelled by multiple annotators, resulting in an inter-annotator agreement of 91.23% on the presence of hand washing and 90.06% on movement classification when washing was detected.

2.1.2. METC dataset

The METC dataset was collected at the Medical Education Technology Centre of Riga Stradins University as a part of a user feedback study [34] involving 72 healthcare specialists. Data collection took place in a single location. Each participant performed three hand-washing trials under varying feedback conditions: no guidance, semi-guidance from a smartphone app, and full guidance from the same app.

Hand-washing sessions were recorded and labelled in real time by a human operator. The dataset contains 212 videos with 212 annotations, covering 13 870 seconds of footage.

2.2 Initial experiments on PSKUS and METC datasets

In the initial experiments, I used MobileNetV2 CNN models. The first experiment involved freezing the base model and training the added layers – a fully connected layer with 128 neurons, and an output layer with 7 neurons – for 30 epochs. The model trained on the PSKUS dataset achieved an F_1 score of 16.72%, while the model trained on the METC dataset performed much better, achieving an F_1 score of 49.74%.

In the second experiment, I unfroze all the model layers and continued training for 30 additional epochs with a reduced learning rate. The results for the PSKUS dataset showed no improvement, with an F_1 score of 15.52%, while the results for the METC dataset demonstrated a substantial improvement, with an F_1 score increasing to 63.89%.

The difference in performance between the datasets highlights the challenge of generalizing across locations and less consistent labelling provided by multiple annotators in the PSKUS dataset. In contrast, experiments

on the METC dataset, with its single annotator and more controlled environment, yielded better results.

2.3 Cross-dataset study of CNN performance

As CNN models in the initial experiments did not demonstrate as high results on the hand-washing movement classification task as those reported in the literature, my collaborators and I conducted a study to investigate whether CNN architectures that perform well on simpler, smaller datasets can generalize to larger, more complex ones. In addition to the PSKUS and METC datasets, we used the publicly available part of the Kaggle hand-washing dataset [35], which consists of 25 scripted hand-washing videos. In our experiments, we maintained focus on lightweight classifiers such as MobileNets, which do not require high-end hardware and therefore are suitable for real-world hand-washing monitoring systems. Additionally, we included multi-stream network architectures and recurrent elements, commonly used in the literature.

2.3.1. Methodology

The baseline model used in the experiments had the same architecture as the model used in the initial experiments. Two additional types of architecture were tested to improve temporal recognition of hand-washing movements. The first was a two-stream network, where one MobileNetV2 model processed RGB input, and the other processed optical flow to capture motion between frames. The second architecture was a recurrent CNN consisting of five parallel MobileNetV2 models combined by a single `Gated Recurrent Unit` (GRU; [36]) module with 256 neurons.

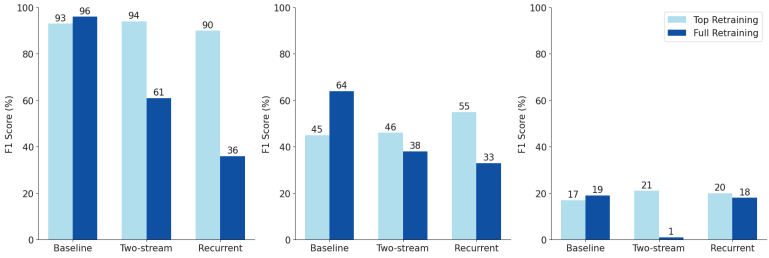
The baseline and recurrent models were used both with the configuration described above and with two additional fully connected layers with 128 neurons in each, added before the output layer.

The models were trained for 20 epochs. We conducted experiments both with retraining the whole model or only its top layers. Additional transfer learning experiments were conducted for the overall duration of 10 epochs to assess model generalization, testing knowledge transfer from simpler datasets to the more complex ones.

2.3.2. Results

The results of the main experiments (Fig. 2.1) showed that the baseline single-frame model performed best on both Kaggle (96% F_1 score) and METC (64% F_1) datasets. Surprisingly, more complex models – the two-stream network and recurrent CNN – demonstrated worse performance.

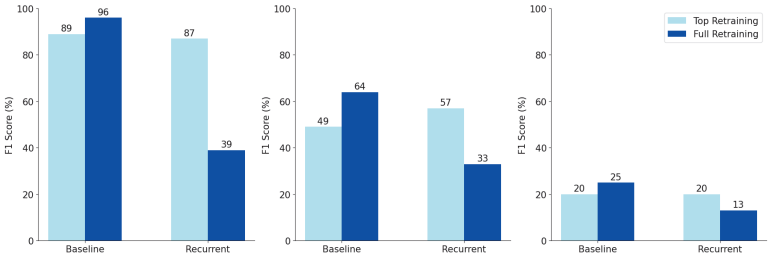
On the PSKUS dataset, the best model (two-stream network with the top retrained) achieved only a 21% F_1, indicating poor generalization to complex real-world data.



(a) Kaggle (b) METC (c) PSKUS

Figure 2.1: F_1 scores of different CNN architectures.

Experiments with adding two fully connected layers (Fig. 2.2) to baseline and recurrent models[1] revealed similar trends: the single-frame model still performed best on Kaggle (96% F_1) and METC (64% F_1) datasets, but improvements on the PSKUS dataset were minimal, with the best F_1 score reaching only 25%.

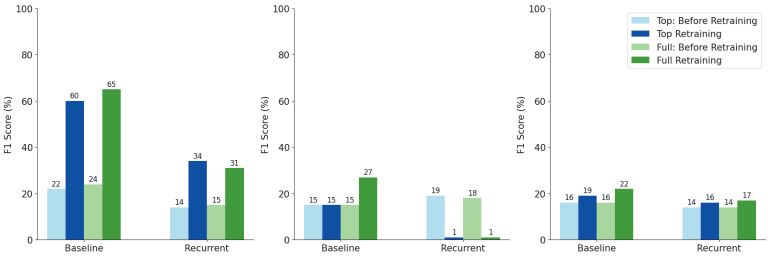


(a) Kaggle (b) METC (c) PSKUS

Figure 2.2: F_1 scores of different CNN architectures with two added fully connected layers.

Finally, transfer learning experiments (Fig. 2.3) showed limited success. The best-performing model, baseline Kaggle-to-METC, achieved 65% F_1 on METC, but generalization to PSKUS remained poor, with a top F_1 of just 27%. This suggests that knowledge learned on simpler datasets does not transfer effectively to more complex real-world datasets, such as PSKUS.

[1] As the two-stream network did not show a major improvement over the two other approaches in any experiment, we excluded this architecture from further experiments.



(a) Kaggle to METC (b) Kaggle to PSKUS (c) METC to PSKUS

Figure 2.3: F_1 scores of different CNN architectures: transfer learning.

Overall, models that performed well on the simpler Kaggle dataset struggled with the more complex, real-world datasets – METC and especially PSKUS. Adding temporal information or extra layers did not improve performance, and full retraining was only beneficial for single-frame models. The findings of the study confirm **the main hypothesis** for all experiments except those on the PSKUS dataset and highlight the importance of considering dataset complexity when evaluating hand-washing classification accuracy in real-world conditions.

3. Semantic Segmentation of Street Views

The task of semantic segmentation of street views is essential for the design of self-driving cars – a promising emerging technology and a lively area of academic and industrial research. Similar to other visual domains, the best results in semantic segmentation of street views have been achieved using CNNs – for instance, models from the state-of-the-art DeepLab library [37]. However, since acquiring and annotating urban images for training CNNs is costly and time-consuming, publicly available datasets of street views labelled for segmentation, e.g., CamVid [38], KITTI [39], and Cityscapes [40], tend to be relatively small.

The use of synthetic data can help address the data scarcity issue. Prior work has demonstrated the utility of synthetic images acquired in video games [41] and custom-made synthetic datasets of street scenes [42]. Arguably, a more accessible alternative is to use open-source driving simulators such as TORCS [43] or CARLA [44]. However, the question remains whether synthetic data generated with such tools are of sufficiently high quality to be useful for training semantic segmentation models.

The goal of the research reported in the present chapter was to improve the accuracy of semantic segmentation of street views by augmenting a dataset of real-world images with synthetic data. In particular, I investigated whether it is possible to improve the accuracy of semantic segmentation by using synthetic data generated with CARLA, which can be done in a relatively simple, fast, and largely automated manner. **The main hypothesis** of the study was that augmenting real-world data with synthetic data would result in the improved accuracy of semantic segmentation of street views.

3.1 Datasets

I used three datasets in my experiments: the real-world Cityscapes dataset, the MICC-SRI dataset [45], and the CCM (Cityscapes-CARLA Mixed) dataset, which I created for this study.

Cityscapes features 5 000 images with fine (pixel-level) annotation; the resolution of images is 1024×2048 pixels, and they were taken in 50 different European cities. For my study, I created a custom dataset split – 2 685 images for training, 290 for validation, and 500 for testing – since the original test set is withheld from open access for benchmarking.

MICC-SRI consists of 11 913 synthetic images generated with CARLA with a resolution of 600×800 pixels, and their segmentation masks. The images have a low degree of photorealism.

My custom-made CCM dataset consists of 2 685 Cityscapes images as well as 46 935 synthetic images that I generated with CARLA. The resolution of synthetic images is 1024×2048 pixels, matching the resolution of Cityscapes images. To generate the data, I ran simulations on several maps available in the then-latest stable release of CARLA (v0.9.12), populating each map with 100 vehicles and 200 pedestrians.

3.2 Methodology

Prior to training CNNs, I preprocessed the data by relabelling semantic segmentation masks and resizing images to ensure compatibility between the datasets. Furthermore, I created several splits of the CCM dataset – CCM-100, CCM-50, and CCM-25 – by augmenting real-world Cityscapes data with a respective percentage of available synthetic data.

After data preprocessing, I conducted semantic segmentation experiments using two CNN models from the DeepLabv3 library – MobileNetV2 and Xception-65. The models were trained using their default settings in the DeepLabv3 library. I used only real-world Cityscapes images for validation and testing the models.

For experiments on MICC-SRI and Cityscapes, the MobileNetV2 models were trained for 1200 epochs on each dataset, and the Xception-65 models were trained for 300 epochs on each dataset. For experiments on CCM and Cityscapes, the MobileNetV2 and Xception-65 models were trained for 200 epochs on Cityscapes and the CCM-100, CCM-50, and CCM-25 dataset splits. Training was conducted on the servers of Riga Technical University (RTU) High Performance Computing (HPC) Center and took $\approx$ 2900 hours of computing in total.

3.3 Results

I report the main results using the standard metrics for semantic segmentation: per-class Intersection over Union (IoU) and mean Intersection over Union (mIoU). Similar to other authors (cf. e.g. [40]), I report and include in the calculations only semantically meaningful classes, excluding such classes as `Other` or `None`.

Results on Cityscapes and MICC-SRI datasets

The results of training MobileNetV2 and Xception-65 on Cityscapes and MICC-SRI are summarised in Tables 3.1 and 3.2.

Table 3.1: Comparison of the accuracy (IoU) of semantic segmentation: MobileNetV2 trained on Cityscapes vs. MobileNetV2 trained on Cityscapes augmented with MICC-SRI.

Class	Cityscapes	Cityscapes & MICC-SRI
Road	**92.66**	92.62
Sidewalk	**67.02**	66.61
Building	**86.48**	86.18
Fences and Walls	**44.46**	43.21
Poles and traffic signs	**57.07**	56.72
Vegetation	**89.52**	89.45
Pedestrians	**76.59**	76.54
Vehicles	**89.65**	89.55
Mean IoU	**75.43**	75.11

Table 3.2: Comparison of the accuracy (IoU) of semantic segmentation: Xception-65 trained on Cityscapes vs. Xception-65 trained on Cityscapes augmented with MICC-SRI.

Class	Cityscapes	Cityscapes & MICC-SRI
Road	**93.69**	93.60
Sidewalk	71.78	**72.70**
Building	**88.67**	88.30
Fences and Walls	**52.20**	49.16
Poles and traffic signs	**63.58**	62.52
Vegetation	**90.75**	90.58
Pedestrians	**81.75**	81.39
Vehicles	**92.29**	92.24
Mean IoU	**79.34**	78.81

As can be seen, the augmentation of Cityscapes with MICC-SRI images did not improve the accuracy of semantic segmentation: in fact, both MobileNetV2 and Xception-65 models performed slightly better when trained only on real-world images than on the augmented dataset. The likely explanation for the worse performance of the models trained on the augmented dataset is the low photorealism of images in the MICC-SRI dataset.

Results on Cityscapes and CCM datasets

The results of training MobileNetV2 and Xception-65 on Cityscapes and the three splits of CCM – CCM-100, CCM-50, and CCM-25 – are reported in Tables 3.3 and 3.4. As can be seen, for both CNN architectures, aug-

mentation with synthetic data improved semantic segmentation accuracy, thus confirming **the main hypothesis**.

Table 3.3: Comparison of the accuracy (IoU) of semantic segmentation: MobileNetV2 trained on the Cityscapes, CCM-100, CCM-50, and CCM-25 datasets.

Class	**Cityscapes**	**CCM-100**	**CCM-50**	**CCM-25**
Building	75.54	79.39	**80.17**	79.18
Fence	00.02	21.49	**24.47**	17.82
Pedestrian & Rider	69.23	67.94	68.92	**69.38**
Pole	10.48	38.51	**38.54**	36.81
Road	88.57	88.46	**89.72**	89.15
Sidewalk	54.51	57.24	**59.79**	58.62
Vegetation	83.90	86.14	**86.41**	85.54
Vehicles	82.20	82.72	82.72	**82.78**
Wall	0.00	19.90	**23.81**	15.95
Traffic Sign	0.00	**35.42**	34.10	25.30
Sky	82.80	85.32	85.83	**85.85**
Traffic light	0.00	**21.32**	15.38	00.13
Water & Terrain	33.61	37.50	**38.73**	38.20
Mean IoU	44.68	55.49	**56.05**	52.67

Table 3.4: Comparison of the accuracy (IoU) of semantic segmentation: Xception-65 trained on the Cityscapes, CCM-100, CCM-50, and CCM-25 datasets.

Class	**Cityscapes**	**CCM-100**	**CCM-50**	**CCM-25**
Building	84.94	85.10	85.08	**85.62**
Fence	37.20	40.19	40.19	**43.44**
Pedestrian & Rider	**78.08**	76.42	76.94	77.92
Pole	45.08	48.50	48.75	**49.26**
Road	**92.31**	91.82	91.43	91.85
Sidewalk	65.80	67.21	67.09	**69.88**
Vegetation	87.71	87.00	87.59	**87.85**
Vehicles	**89.86**	88.82	89.63	89.63
Wall	23.29	28.88	27.69	**31.63**
Traffic Sign	44.42	50.89	55.83	**56.14**
Sky	85.13	88.79	89.70	**90.34**

Table 3.4 - continued

`Traffic Light`	0.00	43.64	42.19	**44.13**
`Water & Terrain`	**46.86**	39.58	35.62	44.22
Mean IoU	57.25	63.14	63.87	**64.46**

Note that the best-performing MobileNetV2 and Xception-65 models were not the ones trained on the CCM splits with the largest amounts of synthetic data: the best-performing MobileNetV2 was trained on CCM-50, while the best-performing Xception-65 was trained on CCM-25. This suggests that using larger amounts of synthetic data for augmentation does not necessarily lead to better performance than augmentation with smaller amounts of such data. Furthermore, this research also demonstrated that setting up a pipeline for generating synthetic data does not have to be costly or difficult, as CARLA allowed for generating a large amount of synthetic data quickly and without much modification to the out-of-the-box installation.

4. Object Detection for a Bin-Picking Task

Single-stage object detectors such as YOLO [46] are often deployed in industrial robotic systems for object grasping, because they offer both high accuracy of object detection and fast inference speed. However, large datasets are needed for training such models. A promising solution to the problem of the availability of data for the development of perceptual systems in robotics is the use of simulations and synthetic data. However, models trained solely on synthetic data often suffer from performance drops when applied to real-world tasks due to the 'reality gap' [47] between artificial and real-world domains. To improve the realism of synthetic data, one can use domain adaptation techniques such as generative adversarial networks (GANs; [48]).

The main goal of the study reported in this chapter was to improve the accuracy of detecting high-visibility (i.e., those on the top of the pile) plastic bottles with YOLOv5 object detector [17] to enable the robotic arm to better grasp them. **The main hypothesis** of the study was that enhancing the photorealism of synthetic images of plastic bottles with GANs before training YOLOv5 on them would result in higher object detection accuracy than using unmodified synthetic images for training.

4.1 Datasets

A number of datasets were used in the study, namely, a real-world bottle image dataset, a synthetic bottle image dataset, and several datasets created by enhancing the synthetic dataset with GANs.

The real-world dataset was acquired in [49] by randomly placing bottles in a plastic container and capturing images with varying bottle positions, camera exposure times, and lighting conditions. The dataset includes 2 060 manually labelled images, 1 760 of which were used for training GANs, while 300 were reserved for testing object detectors.

The synthetic dataset was generated in [49] using the Blender physics simulation engine. It contains 8 800 high-resolution images of bottles randomly dropped into a simulated box, with variations in lighting intensity and viewing angles. Bottles with more than 60% visibility were considered graspable.

To create datasets with enhanced photorealism, CycleGAN [50] was used, since it does not require paired real and synthetic images, which simplifies the process. A number of datasets were created by trying different approaches, particularly:

- *Baseline* CycleGAN – by using CycleGAN with default parameters;

- *Augmented* CycleGAN – by cropping the centre of the image and adding data preprocessing functions such as random contrast, brightness, hue, and saturation to reduce overfitting;
- *Augmented noise* CycleGAN – by adding Gaussian noise to the input to the discriminator to make it more difficult for the discriminator to evaluate images and thus allowing the generator to train longer;
- *Resized convolution* CycleGAN – by replacing the transposed convolution [51] operations in the GAN decoder with resizing followed by convolution to remove checkerboard artifacts [52];
- *Resized transpose* CycleGAN – by applying resized convolution on all layers except the last one, where the transposed convolution was retained. That was done to remove the blur that appeared in the *Resized convolution* CycleGAN dataset images.

After evaluating GAN-enhanced datasets by means of both visual inspection and calculating the Frechet Inception Distance (FID) scores [53], *Augmented noise* CycleGAN and *Resized transpose* CycleGAN datasets were chosen for object detection experiments. These datasets were used with a resolution of 1024×768 pixels, as images with such resolution were found to have a better FID score than images with the smaller resolution of 256×256 pixels used originally.

4.2 Methodology

I conducted object detection experiments using YOLOv5 from the Ultralytics library. The experiments were performed on three datasets: the original synthetic images, the *Augmented noise* CycleGAN 1024×768 dataset, and the *Resized transpose* CycleGAN 1024×768 dataset. I used YOLOv5 models of three different sizes – Small (7.2M parameters), Medium (21.2M parameters), and Extra Large (86.7M parameters). The models were trained for 300 epochs and subsequently tested on a set of 300 real-world images.

4.3 Results

I report the results of the object detection experiments in Table 4.1, using standard metrics: precision, recall, and mean average precision (mAP) for bounding boxes. mAP was calculated for an IoU threshold of 0.5 as well as averaged across IoU thresholds from 0.5 to 0.95 in steps of 0.05.

Table 4.1: Results of the object detection experiments with YOLOv5 on the original synthetic data and synthetic data enhanced with CycleGAN.

Model	**Dataset**	**Precision**	**Recall**	**mAP (threshold 0.5)**	**mAP (avg for IoU $\in$ [0.5 : 0.05 : 0.95])**
YOLOv5 Small	Original synthetic	68.9	89.4	74.2	44.7
	Resized transpose 1024×768	61.5	83.8	63.2	24.5
	Augmented noise 1024×768	**73.1**	**90.5**	**77.7**	**48.0**
YOLOv5 Medium	Original synthetic	69.3	80.2	72.0	42.0
	Resized transpose 1024×768	58.1	78.2	58.9	20.4
	Augmented noise 1024×768	**71.2**	**91.0**	**75.3**	**45.5**
YOLOv5 Extra Large	Original synthetic	71.8	78.5	75.0	41.3
	Resized transpose 1024×768	68.3	86.7	71.5	33.9
	Augmented noise 1024×768	**72.1**	**87.2**	**76.1**	**46.1**

While the models trained on the *Resized transpose* 1024×768 CycleGAN dataset generally performed worse than the models trained on the original synthetic data, models trained on the *Augmented noise* 1024×768 CycleGAN dataset consistently outperformed the models trained on the original synthetic data across all metrics, thus confirming **the main hypothesis**. Note that larger models did not always outperform smaller ones, which suggests that the larger models may have overfitted to the relatively small training datasets.

5. Image Classification for Monitoring the Growth of Organs-on-a-Chip

OOC is a promising biomedical technology that combines tissue engineering and microfluidics to imitate key aspects of human physiology, with the aim of recreating the environment of particular human organs *in vitro*. Currently, monitoring of OOC setups is done by humans, but it would be helpful to automate this process. As one of the key aspects of monitoring is assessing the quality of OOC cell samples, **the goal** of the research reported in this chapter was to develop a classifier for OOC microscopy images. Taking into account the success of CNNs in many biomedical tasks [54, 55, 56], I used the EfficientNet-B7 and MobileNetV3Large CNNs for this purpose and proposed the following hypothesis:

> **Hypothesis 1:** A CNN-based classifier achieves better accuracy on the real-world microscopy OOC image dataset than a putative 'naive' classifier.

The major challenge for the development of DNN-based biomedical image classifiers is the availability of data, because biomedical datasets are often small and imbalanced. A promising solution to this problem is the use of synthetic data. In the research reported in this chapter, I generated synthetic data with Stable Diffusion [57], a latent diffusion model that has demonstrated impressive capabilities for creating different types of imagery. The hypothesis for the experiments with synthetic images was as follows:

> **Hypothesis 2:** The classification accuracy on the real-world microscopy OOC image dataset improves when a CNN-based classifier is trained on the dataset augmented with synthetic data generated with the Stable Diffusion model rather than solely on the real-world image dataset.

5.1 Experiments on the initial OOC image dataset

5.1.1. Data and methodology

The initial OOC image dataset consisted of 822 images of cells belonging to different cell lines: Caco-2 and HUVEC modelling the gut, A549 and HPMEC modelling lung cancer, and HSAEC modelling the lung. The ground truth labelling of images was done by experienced cell biologists; as a result, based on cell morphology and density, images were labelled

into three classes: 'good' (500 images), 'acceptable' (212 images), and 'bad' (110 images).

To augment the initial OOC image dataset, I fine-tuned Stable Diffusion using the low-rank adaptation (LoRA; [58]), an efficient method requiring just several dozen images, and generated two synthetic datasets, with LoRA weights set to 1.0 and 0.8, respectively.

In the initial experiments, I used the EfficientNet-B7 CNN model. During training, the weights of the base model were frozen, whereas two added layers – the `Batch Normalisation` layer and the output layer with 3 neurons – were trained for 30 epochs.

5.1.2. Results

The results of the experiments on the initial dataset are shown in Table 5.1.

Table 5.1: The accuracy of EfficientNet-B7 trained on the initial dataset augmented with synthetic data generated with the Stable Diffusion model fine-tuned with LoRA.

Dataset	**LoRA weight 1.0**	**LoRA weight 0.8**
Synth only (100%)	61.4	62.1
Real & 100% synth	69.9	70.1
Real & 75% synth	69.6	70.7
Real & 50% synth	71.0	69.3
Real & 25% synth	70.7	70.4
Real & 10% synth	72.1	71.8
Baseline (real only)	**72.9**	

The baseline model achieved an accuracy of 72.9%, better than a putative 'naive' classifier (60.8%). However, augmenting the dataset with synthetic images caused a drop in accuracy, with larger amounts of synthetic data leading to poorer performance.

5.2 Experiments on the final OOC image dataset

5.2.1. Data and methodology

The final dataset of OOC images [59] consists of 3 072 images, incorporating the initial dataset. It also features two major differences: first, based on the recommendation of the biology experts, the three-class labelling was replaced by a more straightforward binary classification: 'good' and 'bad'; second, it includes an additional cell line – NHBE bronchial epithelial cells.

The distribution of the images by classes is as follows: 'good' – 1 727 images, 'bad' – 1345 images.

To augment the final OOC image dataset with synthetic data, several methods were employed, namely:

- **image-to-image translation**, which implies such a transformation of an input image into an output image that the latter both retains some features of the former and acquires some new features;
- **inpainting with masks**, which involves selectively modifying an input image by applying a mask to designate which parts of the image should be altered;
- **image interpolation**, creating an intermediate image between two input images;
- **fine-tuning with LoRA**.

Classification experiments were conducted with EfficientNet-B7 and MobileNetV3Large. Training EfficientNet-B7 training involved two stages: first, the model was trained for 30 epochs with the weights of the base model frozen; second, the top 20 layers were fine-tuned for additional 30 epochs. Training MobileNetV3Large followed a similar approach, with the difference being that during the second stage, the top 15 rather than 20 layers were fine-tuned.

5.2.2. Results

The results of the experiments on the final dataset are provided in Table 5.2. After training on the real-world dataset without synthetic augmentation, EfficientNet-B7 achieved an accuracy of 83%, whereas MobileNetV3Large achieved an accuracy of 81%. Both results are better than the accuracy of a putative 'naive' classifier, which is equivalent to the size of the largest class – 56%; therefore, experiments confirmed **Hypothesis 1**.

The results of experiments involving augmentation with synthetic images demonstrated that augmentation with the data generated with image-to-image translation can improve the accuracy of EfficientNet-B7 up to 85% (i.e., by 2%), whereas in case of MobileNetV3Large, several methods – image-to-image translation, inpainting with masks, and interpolation – can improve the accuracy of that model up to 82% (i.e., by 1%). Therefore, experiments confirmed **Hypothesis 2**.

Table 5.2: The accuracy of EfficientNet-B7 and MobileNetV3Large trained on the final dataset dataset augmented with synthetic data generated by image-to-image translation (img2img), inpaiting with masks (inpaint), interpolation (interpol), and the Stable Diffusion model fine-tuned with LoRA.

Model	**Dataset**	**img2img**	**inpaint**	**interpol**	**LoRA**
EffNet-B7	Real & 100% synth	**85**	84	80	82
	Real & 75% synth	82	83	81	78
	Real & 50% synth	83	82	80	83
	Real & 25% synth	83	80	81	81
	Baseline (real only)		83		
MobNetV3L	Real & 100% synth	81	**82**	79	76
	Real & 75% synth	81	**82**	79	77
	Real & 50% synth	81	81	80	80
	Real & 25% synth	**82**	80	**82**	80
	Baseline (real only)		81		

Conclusion

This PhD thesis was focused on the applications of computer vision methods to three major image understanding tasks: image classification, object detection, and semantic segmentation. Specifically, I used computer vision methods for solving the following real-world problems:

- classification of hand-washing movements in a clinical setting to automate the monitoring of compliance by medical personnel with hand hygiene standards (Chapter 2);
- semantic segmentation of street views to enhance perception modules of self-driving cars (Chapter 3);
- detection of plastic bottles that can be picked up by a robotic arm to automate the production line in a manufacturing facility (Chapter 4);
- classification of microscopy images to automate the monitoring of the growth of organs-on-a-chip (Chapter 5).

I adopted CNNs for these tasks, leveraging models pretrained on large datasets such as ImageNet [60] and MS COCO [61] and making use of transfer learning and fine-tuning to adapt them to much smaller datasets that I was working with. For some of the tasks, I was also able to utilise suitable publicly available datasets: the Cityscapes dataset was essential for my work on improving semantic segmentation of street views, while the Kaggle dataset allowed me to establish a baseline for the assessment of performance of CNN-based hand-washing movement classifiers. Furthermore, I leveraged open-source assets such as CARLA and Stable Diffusion for generating synthetic data, which I could then use for augmenting real-world datasets. However, data acquisition and labelling was still one of the major challenges, particularly when working on hand-washing movement recognition task, as there were no publicly available datasets for that task of sufficient size, and when working on designing an OOC image classifier, as the amount of available data was rather small.

Despite the challenges that I outlined above, I consider the results of the studies that laid the foundation of this thesis to be rather successful, in particular:

- In the research on hand-washing movement classification, excellent results – an F_1 score of 96% – were achieved on the Kaggle dataset and satisfactory results – an F_1 score of 64% - were achieved on the METC dataset. While none of the models achieved good performance on the complex and noisy PSKUS dataset, the dataset in question is a valuable asset for further studies on hand-washing movement classification.

- In the research on improving the accuracy of semantic segmentation of street views, augmentation of Cityscapes with CARLA-generated data resulted in higher accuracy of the MobileNetV2 and Xception models than the accuracy of the models trained only on Cityscapes.
- In the research on detecting graspable bottles, the state-of-the-art object detector YOLOv5 demonstrated that it can efficiently – the best model achieved a mAP of 77.7% with a threshold of 0.5 – detect bottles with the visibility above a certain threshold in a pile of similarly looking objects.
- In the research on OOC microscopy image classification, the best EfficientNet-B7 model achieved an accuracy of 85%, while the best MobileNetV3Large model achieved an accuracy of 82%. Since these best-performing models were trained on the dataset augmented with images generated with Stable Diffusion, this study also contributed to the emerging research of generating biomedical imagery with large generative models for training DNNs.

These results demonstrate that **the goal of the thesis** – to provide efficient solutions for applied image understanding tasks – was achieved for all tasks except the classification experiments on the PSKUS dataset. **The scientific novelty** of the thesis lies in providing effective solutions for novel datasets, where no prior solutions existed, as well as offering improved methods for already explored datasets, leveraging CNNs on both real-world and synthetic data.

The results of the studies presented in this thesis have been published in six scholarly articles (with at least two more publications forthcoming) indexed in Elsevier Scopus and/or Web of Science database and two scholarly publications not indexed in these databases as well as presented at four conferences. However, there is still ample room for further work. The primary objective for the near future is to apply the insights and knowledge gained during the work on this thesis to solve image understanding tasks in other ongoing research projects. The major goal for the more distant yet, I believe, still foreseeable future is to contribute to the development of models truly capable of image *understanding*.

Bibliography

[1] B. Jähne, ed., *Computer Vision and Applications: A Guide for Students and Practitioners*. Elsevier, 2000.

[2] I. Pitas, *Digital Image processing Algorithms and Applications*. John Wiley & Sons, 2000.

[3] Y.-J. Zhang and Y.-J. Zhang, "Image engineering," in *Handbook of Image Engineering*, pp. 55–83, Springer, 2021.

[4] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. Springer Science & Business Media, 2012.

[5] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[6] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[7] N. O'Mahony *et al.*, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1*, pp. 128–144, Springer, 2020.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[9] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[10] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[13] A. Howard *et al.*, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[14] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.

[17] G. Jocher, "YOLOv5 by Ultralytics." [Online], 2020. Available: `https://github.com/ultralytics/yolov5`. Accessed 7 August 2024.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[19] A. Glassner, *Deep learning: A visual approach*. No Starch Press, 2021.

[20] S. Priya and R. A. Uthra, "Deep learning framework for handling concept drift and class imbalanced complex decision-making on streaming data," *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 3499–3515, 2023.

[21] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[22] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[23] S. L. Barnes, D. J. Morgan, A. D. Harris, P. C. Carling, and K. A. Thom, "Preventing the transmission of multidrug-resistant organisms: Modeling the relative importance of hand hygiene and environmental cleaning interventions," *Infection Control & Hospital Epidemiology*, vol. 35, no. 9, pp. 1156–1162, 2014.

[24] K. J. Mckay, R. Z. Shaban, and P. Ferguson, "Hand hygiene compliance monitoring: Do video-based technologies offer opportunities for the future?," *Infection, Disease & Health*, vol. 25, no. 2, pp. 92–100, 2020.

[25] World Health Organization, *WHO guidelines on hand hygiene in health care*. World Health Organization, 2009.

[26] D. J. Gould, D. Moralejo, N. Drey, J. H. Chudleigh, and M. Taljaard, "Interventions to improve hand hygiene compliance in patient care," *Cochrane database of systematic reviews*, no. 9, 2017.

[27] N. Masroor, M. Doll, M. Stevens, and G. Bearman, "Approaches to hand hygiene monitoring: From low to high technology approaches," *International Journal of Infectious Diseases*, vol. 65, pp. 101–104, 2017.

[28] B. C. Knepper, A. M. Miller, and H. L. Young, "Impact of an automated hand hygiene monitoring system combined with a performance improvement intervention on hospital-acquired infections," *Infection Control & Hospital Epidemiology*, vol. 41, no. 8, pp. 931–937, 2020.

[29] R. T. Ellison III, C. M. Barysauskas, E. A. Rundensteiner, D. Wang, and B. Barton, "A prospective controlled trial of an electronic hand hygiene reminder system," in *Open Forum Infectious Diseases*, vol. 2, p. ofv121, Oxford University Press, 2015.

[30] J. M. Boyce *et al.*, "Impact of an automated hand hygiene monitoring system and additional promotional activities on hand hygiene performance rates and healthcare-associated infections," *Infection Control & Hospital Epidemiology*, vol. 40, no. 7, pp. 741–747, 2019.

[31] S. Yeung *et al.*, "Vision-based hand hygiene monitoring in hospitals," *AMIA*, 2016.

[32] G. Li *et al.*, "Hand gesture recognition based on convolution neural network," *Cluster Computing*, vol. 22, no. 2, pp. 2719–2729, 2019.

[33] A. Nagaraj, M. Sood, C. Sureka, and G. Srinivasa, "Real-time action recognition for fine-grained actions and the hand wash dataset," *arXiv:2210.07400*, 2022.

[34] O. Zemlanuhina *et al.*, "Influence of different types of real-time feedback on hand washing quality assessed with neural networks/simulated neural networks," in *SHS Web of Conferences*, vol. 131, pp. 1–13, EDP Sciences, 2022.

[35] "Sample: Kaggle Hand Wash Dataset." [Online], 2019. Available: https://www.kaggle.com/realtimear/hand-wash-dataset. Accessed 18 February 2024.

[36] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv:1409.1259*, 2014.

[37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *arXiv:1412.7062*, 2014.

[38] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[39] H. Abu Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, vol. 126, pp. 961–972, 2018.

[40] M. Cordts *et al.*, "The Cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[41] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*, pp. 102–118, Springer, 2016.

[42] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. Van Gool, "Semantic understanding of foggy scenes with purely synthetic data," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 3675–3681, IEEE, 2019.

[43] B. Wymann *et al.*, "TORCS: The open racing car simulator." [Online], 2015. Available: https://www.cse.chalmers.se/~chrdimi/papers/torcs.pdf. Accessed 29 June 2024.

[44] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Conference on robot learning*, pp. 1–16, PMLR, 2017.

[45] L. Berlincioni, F. Becattini, L. Galteri, L. Seidenari, and A. Del Bimbo, "Road layout understanding by generative adversarial inpainting," in *Inpainting and Denoising Challenges*, pp. 111–128, Springer, 2019.

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

[47] N. Jakobi, P. Husbands, and I. Harvey, "Noise and the reality gap: The use of simulation in evolutionary robotics," in *Advances in Artificial Life: Third European Conference on Artificial Life*, pp. 704–720, Springer, 1995.

[48] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[49] J. Arents *et al.*, "Synthetic data of randomly piled, similar objects for deep learning-based object detection," in *International Conference on Image Analysis and Processing*, pp. 706–717, Springer, 2022.

[50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[51] W. Shi *et al.*, "Is the deconvolution layer the same as a convolutional layer?," *arXiv:1609.07009*, 2016.

[52] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, 2016.

[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[54] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, F. A. Oluwaseyi, and K. I. Okesola, "Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches," *International Journal of Advanced Computer Research*, vol. 11, no. 53, p. 51, 2021.

[55] A. Iqbal, M. Sharif, M. A. Khan, W. Nisar, and M. Alhaisoni, "FF-UNet: a U-shaped deep convolutional neural network for multimodal biomedical image segmentation," *Cognitive Computation*, vol. 14, no. 4, pp. 1287–1302, 2022.

[56] S. Sharma *et al.*, "Performance evaluation of the deep learning based convolutional neural network approach for the recognition of chest X-ray images," *Frontiers in oncology*, vol. 12, 2022.

[57] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[58] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.

[59] V. Movčana *et al.*, "Organ-on-a-chip (OOC) image dataset for machine learning and tissue model evaluation," *Data*, vol. 9, no. 2, p. 28, 2024.

[60] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, IEEE, 2009.

[61] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision–ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, vol. 8693, pp. 740–755, Springer, 2014.

Acknowledgements

I would like to express my deepest gratitude to a number of people whose support and encouragement have been pivotal in bringing this thesis to completion.

First and foremost, I would like to thank my thesis advisor, Dr. Roberts Kadiķis, for his invaluable guidance, expertise, and patience. His insights and advice have been fundamental in shaping my research career in general as well as this thesis in particular.

I am immensely grateful to my family for their unwavering support throughout this journey: to my wife Ilze, for her love, patience and companionship; to my daughters Laura, Alise, Adriāna, Emīlija, and Karolīna, who fill my life with joy and inspiration; to my mother Ņina, for her belief in my success.

A special thank you goes to my colleagues at the Institute of Electronics and Computer Science (EDI), especially Dr. Modris Greitāns and Dr. Kaspars Ozols for creating a supportive research environment, and Dr. Atis Elsts, Dr. Jānis Judvaitis, and fellow PhD candidates Krišjānis Nesenbergs, Didzis Lapsa, and Anatolijs Zencovs for the opportunities to engage in stimulating discussions.

I am very grateful to all the co-authors of my publications, with whom I have had the pleasure and honour to collaborate. Their contributions have been instrumental in my research.

I would also like to acknowledge my colleagues at the University of Latvia (UL), especially Professor Juris Borzovs, Professor Zane Bičevska, Professor Jānis Zuters, Professor Laila Niedrīte, and Professor Guntis Arnicāns for providing me with the opportunities to teach at UL and invaluable advice and kind help in various academic and administrative matters. I wish to thank the administrative staff at the UL, particularly Ārija Sproģe, Dace Mileika, and the late Anita Ermuša, for their kind assistance in various organisational matters. I am also very grateful to Professor Jānis Zuters and Associate Professor Edgars Celms for being examiners at my PhD exam, to Professor Uldis Straujums, Professor Guntis Arnicāns, and Professor Juris Borzovs for providing valuable feedback during the public discussion of an earlier version of this thesis, and to Associate Professor Jevgēnijs Vihrovs for kindly sharing with me the LaTeX template of his PhD thesis, thus saving me a lot of time and effort on formatting this work.

I extend my heartfelt thanks to my students at the University of Latvia, who have greatly contributed to my development as an educator, thus also making me a better researcher.

Last but certainly not least, I wish to thank my dogs Pērle, Kara (oh, sweet little Kara!), Lesija, and Bella, and my pet parrots Rons, Lira, Solo, Frodo, and the late Taira for their companionship and the joy they bring into my life. They have been a source of comfort and relief for me during the most stressful times.

This academic journey has been challenging, rewarding, and at times, challenging once more, and I am thankful to have had such a supportive network of family, colleagues, students, and friends. Once again, thank you all!