Synthetic data for visual machine learning: A data-centric approach

Apostolia Tsirikoglou

Linköping studies in science and technology. Dissertation, No. 2202

SYNTHETIC DATA FOR VISUAL MACHINE LEARNING: A DATA-CENTRIC APPROACH

Apostolia Tsirikoglou

Division of Media and Information Technology
Department of Science and Technology
Linköping University, SE-601 74 Norrköping, Sweden
Norrköping, February 2022

Description of the cover image

The cover of this book is a photo mosaic that uses synthetic visual data developed during the thesis. The background/base and tile images are samples from all the enclosed works, including street scene parsing, cancer detection, and weather classification.

Synthetic data for visual machine learning: A data-centric approach

Copyright © 2022 Apostolia Tsirikoglou (unless otherwise noted)

*Division of Media and Information Technology
Department of Science and Technology
Linköping University, Campus Norrköping
SE-601 74 Norrköping, Sweden*

ISBN: 978-91-7929-174-7 (Print)

ISBN: 978-91-7929-175-4 (PDF)

ISSN: 0345-7524

Printed in Sweden by LiU-Tryck, Linköping, 2022

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

<https://creativecommons.org/licenses/by-nc/4.0/>

“ That is what **learning** is.
You suddenly understand something
you've understood all your life,
but in a new way ”.

Doris Lessing

Deep learning allows computers to learn from observations, or else *training data*. Successful application development requires skills in neural network design, adequate computational resources, and a training data distribution that covers the application domain. We are currently witnessing an artificial intelligence (AI) outbreak with enough computational power to train very deep networks and build models that achieve similar or better than human performance. The crucial factor for the algorithms to succeed has proven to be the training data fed to the learning process. Too little or low quality or out-of-the-target distribution data will lead to poorly performing models no matter the capacity and the data regularization methods.

This thesis is a *data-centric* approach to AI and presents a set of contributions related to *synthesizing images* for training supervised visual machine learning. It is motivated by the profound potential of synthetic data in cases of low availability of captured data, expensive acquisition and annotation, and privacy and ethical issues. The presented work aims to generate images similar to samples drawn from the target distribution and evaluate the generated data as the sole training data source and in conjunction with captured imagery. For this, two synthesis methods are explored: *computer graphics* and *generative modeling*. Computer graphics-based generation methods and synthetic datasets for computer vision tasks are thoroughly reviewed. In the same context, a system employing procedural modeling and physically-based rendering is introduced for data generation for urban scene understanding. The scheme is flexible, easily scalable, and produces complex and diverse images with pixel-perfect annotations at no cost. Generative Adversarial Networks (GANs) are also used to generate images for small data scenarios augmentation. The strategy advances the model's performance and robustness. Finally, ensembles of independently trained GANs investigate ways to improve images' diversity and create synthetic data to serve as the only training source.

The application areas of the presented contributions relate to two image modalities, *natural* and *histopathology* images, to cover different aspects in the generation methods and the tasks' characteristics and requirements. There are showcased synthesized examples of natural images for automotive applications and weather classification, and histopathology images for breast cancer and colon adenocarcinoma metastasis detection. This thesis, as a whole, promotes data-centric supervised deep learning development by highlighting the potential of synthetic data as a training data resource. It emphasizes the control over the formation process, the ability of multi-modality formats, and the automatic generation of annotations.

Keywords: training data, synthetic images, computer graphics, generative modeling, natural images, histopathology, digital pathology, machine learning, deep learning

Populärvetenskaplig Sammanfattning

Termen *artificiell intelligens* (AI) används för att beskriva datorimplementerade algoritmer vars uppgift är att efterlikna mänsklig förmåga. Det kan t.ex. vara regelbaserade system för att utföra olika uppgifter, eller för att imitera mänskligt beteende. Ett område inom AI är *maskininlärning*, vilket beskriver metoder för att utifrån data automatiskt lära datorer att utföra specifika uppgifter. Exempel på uppgifter innefattar språkförståelse, ansiktsgenkänning, objektidentifiering, bildgenerering, scenförståelse för autonoma fordon, medicinsk diagnostik, t.ex. cancer-igenkänning, riktad reklam och individanpassade flöden på social media, och många fler. Gemensamt för uppgifterna är att de går att formulera som problem där korrekta resultat lärs in från insamlad data, för att sedan appliceras på ny data.

En framgångsrik typ av maskininlärning är *djupinlärning*, där *artificiella neuronnät* beskriver komplicerade modeller med hjälp av en uppsättning lager som sekventiellt bearbetar, tolkar och transformerar indata. Med moderna tekniker går det att träna mycket djupa nätverk på stora mängder data och bygga modeller som i vissa applikationer uppnår liknande eller bättre prestanda än en människa. En av de avgörande faktorerna för framgångsrik djupinlärning är den träningsdata som används i inlärningsprocessen. För lite data, data av låg kvalitet, eller data som skiljer sig från den data som en modell är tänkt att användas på, kommer att leda till otillräckliga resultat oavsett hur hög kapacitet modellen har och hur effektiv metoden är som används för inlärning.

Inom många bildbaserade tillämpningar av maskininlärning kan det vara svårt, tidskrävande och kostsamt att samla in och annotera tillräckligt med bilder för att träna ett djupt neuronnät till hög prestanda. En lovande metod för att komma till rätta med detta problem är att använda datorgenererade, *syntetiska*, bilder. Dessa kan skapas med hjälp av *datorgrafik*, där virtuella objekt och miljöer formuleras. Genom att simulera hur ljus propagerar genom den virtuella världen och renderar en bild på en virtuell kameran sensor, kan realistiska bilder genereras. En annan metod är att använda *generativ maskininlärning*, där en bild formuleras som utdata från en modell. Denna kan tränas att generera helt nya bilder, eller att modifiera redan existerande bilder.

Denna avhandling fokuserar på *datacentrerade* strategier för maskininlärning och presenterar en uppsättning bidrag relaterade till syntetisering av bilder för användning som träningsdata i övervakad visuell maskininlärning. Arbetet motiveras av den potential syntetisk data har påvisat i situationer då det finns otillräcklig mängd riktig data, eller då känslig data inte kan tillgängliggöras för träning. Det presenterade arbetet syftar till att generera bilder som efterliknar den tilltänkta användningssituationen, och utvärderas både som enda källa till träningsdata, samt som augmenteringsverktyg för att utöka ett befintligt dataset av riktiga bilder. För dessa ändamål utforskas både datorgrafik och generativ modellering. En undersökning presenteras för att belysa vilka metoder

som finns tillgängliga för att med datorgrafik skapa syntetiska datauppsättningar för datorseendeuppgifter. Ett system introduceras som använder procedurell modellering och fysiskt baserad rendering för datagenerering av stadsmiljöer. Systemet är flexibelt, skalbart och producerar komplexa bilder med pixelperfekta annoteringar. I ett annat projekt används Generative Adversarial Networks (GANs) för att utöka mängden bilder. I situationer med begränsad mängd data förbättrar denna strategi en tränad modells prestanda och robusthet. Slutligen undersöks möjligheterna i att kombinera ett flertal tränade GANs för att förbättra mångfalden i ett genererat dataset, vilket möjliggör användning av enbart syntetisk data i träning.

Tillämpningsområdena för de presenterade bidragen kan delas in i två bildmodaliteter, *naturliga* och *histopatologi*-bilder, vilka täcker in olika aspekter i de krav som ställs på genererade bilder och deras användning. Syntetisering av naturliga bilder testas i tillämpningar inom autonoma fordon samt väderklassificering, och syntetiserade histopatologiska bilder används för att förbättra detektion av bröst- och tjocktarmscancer. Avhandlingen i sin helhet främjar ett datacentrerad perspektiv på djupinlärning genom att lyfta fram den potential syntetisk data har att fungera som träningsdataresurs. Avhandlingen betonar vikten av att syntetisera data med kontroll över bildgenereringsprocessen, med förmåga att täcka in multimodala bildformat, och som automatiskt genererar annoteringar.

Acknowledgments

According to the Merriam-Webster dictionary, the definition of *learning* is “*the activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something*”. It is “*the activity of someone who learns*”. Learning is one of the driving forces of the world, and it always includes two parties that participate in the flow of knowledge directly or indirectly, consciously or unconsciously. This thesis is not an exception. Although it deals with machine learning and synthetic data, I learned a lot in dimensions that I did not expect, and every experience and everyone involved was entirely real.

My main and co-supervisors are the people that I owe and want to acknowledge first. Jonas Unger, I learned so much from you. You introduced me to the machine learning and image synthesis world, you guided me through research practices and intriguing concepts, we discussed tons of ideas, you have been my main academic support, and you showed me that in difficult times it really matters to have compassion. Thank you for the ride! It had it all, and it was truly fantastic! Gabriel Eilertsen, you are this brilliant example of researcher and colleague that I would like to become. Your dedication, guidance, and invaluable help were catalysts in this work. Tusen tack! Anders Ynnerman, thank you for making the division of Media and Information Technology the place I consciously wanted to pursue my PhD. Moreover, thank you for the trust that you personally showed me during these years.

I could not have managed to present this work without the co-authors of the papers we have worked on throughout these years. My sincere thanks to all of my collaborators: Joel Kronander, Magnus Wrenninge, Karin Stacke, Martin Lindvall, Claes Lundström, Marcus Gladh, Daniel Sahlin, Gabriel Eilertsen, Jonas Unger, Per Larsson, Tanaboon Tongbuasirilai, Andrew Gardner, Saghi Hajisharif, Param Hanji, and Rafał K. Mantiuk.

Moreover, this thesis has been proofread by Gabriel Eilertsen, Jonas Unger, Panagiotis Tsirikoglou, Theodoros Tsirikoglou, Foivos Sklivaniotis, Saghi Hajisharif, and Tanaboon Tongbuasirilai. Thank you all for helping me to improve it!

The group of Computer Graphics and Image Processing has been a place of academic and personal growth, and genuine fun, all because of my amazing colleagues, old and new! Sincere thanks to all the members of our group! Saghi Hajisharif, few things came effortlessly these years and your friendship is the most important one. Thank you for sharing knowledge, experiences, feelings, and countless moments of pure joy and laughter! I keep them all close to my heart and look forward to more. Ehsan Miandji, the combination of your passion for your field and your kind heart is rare and unique. Thank you for inspiring me and being my friend! Tanaboon Tongbuasirilai, thank you for being the calm force and always willing to help! Karin Stacke, thank you for all the interesting discussions, research-related or not, book recommendations, and the warm smile that you always bring with you!

Eva Skärblom and Carina Enemård, thank you for your support, knowledge, patience, and kindness all of these years! Gun-Britt Löfgren, thank you for the last-minute administrative help with this thesis! I would also like to acknowledge all the colleagues in the division that we have collaborated with in projects out of this thesis or been involved in constructive, or simply fun, discussions!

During this “*strange*” times of a PhD, I am grateful I had the privilege to have my friends and family by my side. Christina, Evi, Mary, and Xenia, thank you for being steadily these friends that I can always find comfort and love for many, many years now. Tilemache, Vicky, Argyri, our company kept me sane and filled me with joy! To all my dear friends in Stockholm and Norrköping, heartfelt thanks! The happy times we spent together were so much needed! Maria, Marko, and Myrto thank you for your constant support and love! Eleni, Yianni, Kosta, and Spiro thank you for making our adulthood pact even better than our childhood one!

Maria and Panagioti, you are the definition of unconditional love and no-matter-what for me. You have helped me in so many ways all these years that I cannot verbalize my gratitude. Mum and dad, Lina and Thodori, you are the best parents I could ever have. Thank you for always loving me, believing in me, and helping me with all your means.

Foivo, nothing would be the same without you, including this thesis. You have been by my side from the beginning to the end. You lifted me every time things were hard, genuinely celebrated every little or bigger success, and have been patient, supporting, and truly caring. I cannot express how thankful I am. Next and with you, I am learning what *deep loving* is. This thesis is for you.

Apostolia Tsirikoglou.
Norrköping, February 2022

This thesis is based on the following publications:

- Paper A** A. Tsirikoglou, G. Eilertsen, and J. Unger. A Survey of Image Synthesis Methods for Visual Machine Learning. *Computer Graphics Forum*, 39(6):426–451, 2020.
- Paper B** A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger. Procedural Modeling and Physically Based Rendering for Synthetic Data Generation in Automotive Applications. *arXiv eprint arXiv:1710.06270*, 2017.
- Paper C** A. Tsirikoglou, K. Stacke, G. Eilertsen, M. Lindvall, and J. Unger. A Study of Deep Learning Colon Cancer Detection in Limited Data Access Scenarios. In *ICLR Workshop on AI for Overcoming Global Disparities in Cancer Care (AI4CC)*. arXiv, 2020.
- Paper D** A. Tsirikoglou, K. Stacke, G. Eilertsen, and J. Unger. Primary Tumor and Inter-Organ Augmentations for Supervised Lymph Node Colon Adenocarcinoma Metastasis Detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 624–633. Springer International Publishing, 2021.
- Paper E** A. Tsirikoglou, M. Gladh, D. Sahlin, G. Eilertsen, and J. Unger. Generative inter-class transformations for imbalanced data weather classification. *London Imaging Meeting*, 2021(1):16–20, 2021.
- Paper F** G. Eilertsen, A. Tsirikoglou, C. Lundström, and J. Unger. Ensembles of GANs for Synthetic Training Data Generation. In *ICLR Workshop on Synthetic Data Generation (SDG)*. arXiv, 2021.

Additional publications that were conducted during this research journey, but not included in this thesis, are listed below:

1. A. Tsirikoglou, J. Kronander, P. Larsson, T. Tongbuasirilai, A. Gardner, and J. Unger. Differential Appearance Editing for Measured BRDFs. In *ACM SIGGRAPH 2016 Talks, SIGGRAPH '16*, New York, NY, USA, 2016. Association for Computing Machinery
2. G. Eilertsen, S. Hajisharif, P. Hanji, A. Tsirikoglou, R. K. Mantiuk, and J. Unger. How to cheat with metrics in single-image HDR reconstruction. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2021

Generally speaking, the success of supervised deep learning applications is measured using two main factors: the available *training data*, and the *network architecture and resources*. This thesis is a data-centric approach to deep learning development by supporting image synthesis as an alternative training data resource. Data generation is explored for two general image categories: natural images, with all of their content complexity and diversity (Papers A, B, E, and F for simple low-resolution examples) used in typical computer vision tasks, and histopathology images that exhibit seemingly repeated patterns due to the cells and tissue structure (Papers C, D, and F) used in digital pathology. For these categories, image synthesis is approached both from a classical computer graphics (Papers A and B) and a generative modeling perspective (Papers C, D, E, and F). Finally, the synthesized imagery is evaluated as an augmentation source to real/captured data (Papers C, D, and E) or for fine-tuning real data (Paper B), as well as the sole source of training data (Papers B and F).

Paper A [A. Tsirikoglou, G. Eilertsen, and J. Unger. A Survey of Image Synthesis Methods for Visual Machine Learning. *Computer Graphics Forum*, 39\(6\):426–451, 2020](#)

This work provides a comprehensive summary of the, at the time, existing image synthesis methods for visual machine learning. It focuses on the computer graphics aspects of the image data and generation process and provides a relevant taxonomy based on modeling and rendering and some hybrid techniques. It considers methods that create stand-alone training sets for machine learning optimization and classifies them according to the computer vision applications they are used in. Finally, each method is assessed in terms of quality and reported performance, providing a hint for its expected learning potential.

This is an individual work and literature study of the author. Its final version was complemented by the co-authors. It was published in the *Computer Graphics Forum* journal [251] and presented as a state-of-the-art report (STAR) at *Eurographics 2021* planned to be held in Vienna, Austria (virtual event due to the Covid-19 pandemic).

Paper B [A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger. Procedural Modeling and Physically Based Rendering for Synthetic Data Generation in Automotive Applications. *arXiv eprint arXiv:1710.06270*, 2017](#)

This paper presents a novel approach for generating highly realistic, annotated synthetic training data for deep learning solved computer vision tasks inspired by the film and visual effects industry. The main contribution is a procedural world modeling approach coupled with physically-based rendering and comes with high levels of content and process flexibility, accuracy, generated image diversity, and pipeline scalability. Its efficacy is showcased for semantic segmentation for urban scene understanding applied in, e.g., autonomous driving, and

the resulting synthetic data are tested as a training set for both with our without fine-tuning on captured images, producing state-of-the-art results at the time.

This work was conducted in collaboration with 7D Labs Inc.. The author was responsible for evaluating the generated synthetic data over their production iterations and participating in results formation, analysis and assessment. The paper was written in a collaborative effort with the co-authors and published on *arXiv* [250].

Paper C A. Tsirikoglou, K. Stacke, G. Eilertsen, M. Lindvall, and J. Unger. A Study of Deep Learning Colon Cancer Detection in Limited Data Access Scenarios. In *ICLR Workshop on AI for Overcoming Global Disparities in Cancer Care (AI4CC)*. arXiv, 2020

Here, a mitigation strategy for limited training data access scenarios in digital histopathology is introduced. It is driven by the scarcity of labeled data for metastasis detection in the lymph nodes and sets the ground to explore whether mutual structure between tissues can be used to develop general techniques, i.e., if data from one cancer type could have diagnostic value for other cancers elsewhere. The idea is exemplified for colon cancer metastasis detection, and as data sources are tested primary colon tumor tissue and cancer data from a different organ (breast). This work is the precursor of Paper D.

The author was responsible for the setup and execution of the experiments and results production. The paper was written together with the co-authors and presented in the *ICLR 2020 workshop on AI for Overcoming Global Disparities in Cancer Care* planned to be held in Addis Ababa, Ethiopia (virtual event due the to Covid-19 pandemic) [252].

Paper D A. Tsirikoglou, K. Stacke, G. Eilertsen, and J. Unger. Primary Tumor and Inter-Organ Augmentations for Supervised Lymph Node Colon Adenocarcinoma Metastasis Detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 624–633. Springer International Publishing, 2021

This paper extends Paper C and proposes training data augmentation alternatives for colon carcinoma metastasis detection when there is limited or no representation of the target domain. Metastasis detection in the lymphatic system is especially challenging due to the tissue’s low tumor-to-non-tumor ratio, resulting in labor- and time-intensive annotation processes for the pathologists. The paper presents an inter-organ approach utilizing already available annotated tissue from two different organs (breast and skin), and an intra-organ approach, utilizing primary tumor in the colon. In addition to the direct mix strategy, a generative adversarial network (GAN) with cycle consistency (CycleGAN) is used for transforming the proposed augmentations to the target domain. The results show increased models’ performance but also robustness, a factor equally important to the reported accuracies.

The author was responsible for the problem definition specifics, and the setup and execution of the presented large-scale study of cross-validated experiments. The paper was written in collaboration with the co-authors and presented in the

Medical Image Computing and Computer Assisted Intervention 2021 planned to be held in Strasbourg, France (virtual event due to the Covid-19 pandemic) [254].

Paper E A. Tsirikoglou, M. Gladh, D. Sahlin, G. Eilertsen, and J. Unger. Generative inter-class transformations for imbalanced data weather classification. *London Imaging Meeting, 2021*(1):16–20, 2021

This work presents a systematic evaluation of GANs as a tool of training data synthesis for imbalanced small datasets of natural images for single-image weather classification. The synthesized data, built by noise-to-image GANs that use samples of the underrepresented class, and image-to-image translation GANs (CycleGANs) that perform inter-class transformations, augment the captured training sets and are compared to classical augmentations. The results show the potential of GAN-generated inter-class augmentations, opening up the possibilities of this image synthesis method in scenarios where data is complex, diverse, and scarce.

The author set up and performed the experiments, and provided results and insights for their analysis. The weight demodulation modification of the image-to-image translation GAN and the noise-to-image GAN trainings were provided by the rest of the team. The paper was written jointly with the co-authors and presented in the *London Imaging Meeting 2021* (virtual event due to the Covid-19 pandemic) [253].

Paper F G. Eilertsen, A. Tsirikoglou, C. Lundström, and J. Unger. Ensembles of GANs for Synthetic Training Data Generation. In *ICLR Workshop on Synthetic Data Generation (SDG)*. arXiv, 2021

This paper investigates the use of GAN-generated synthetic images as the only source of training data and shows that it is important to use multiple GANs to improve the diversity of the created data. The results reveal that independently trained GANs could be advantageous compared to boosting strategies, with similar or better performance and less prone to over-fitting. The insights are derived by observing the ensemble performance for various kinds of generated data, from simple 2D data to the more realistic scenario of histopathology images for breast cancer detection, where image synthesis could be used in isolation from real data in the name of data scarcity, privacy, and anonymization.

The author contributed to the experimental setup and research formulation of a prestudy which led to this publication. The work was presented in the *ICLR 2021 workshop on Synthetic Data Generation* (virtual event due to the Covid-19 pandemic) [64].

Contents

Abstract	i
Populärvetenskaplig Sammanfattning	iii
Acknowledgments	v
List of Publications	vii
Contributions	ix
1 Introduction	1
1.1 Visual data	2
1.2 Image synthesis	3
1.2.1 Computer graphics	5
1.2.2 Generative image modeling	5
1.3 Deep learning	6
1.3.1 Training data	7
1.4 Objectives	8
1.5 Outline	12
2 Background	15
2.1 Deep learning	16
2.1.1 Neural networks	16
2.1.2 Basic concepts	18
2.1.3 Applications	21
2.2 Computer vision	22
2.3 Digital pathology	23
3 Computer graphics	27
3.1 Modeling	28
3.1.1 Basics	29
3.1.2 Common practices	30
3.1.3 Procedural modeling	31
3.2 Rendering	32
3.2.1 Light transport theory	32
3.2.2 Light transport simulation	35
4 Generative modeling	39
4.1 Fundamentals	40
4.2 Deep generative models	41
4.3 Generative adversarial networks	42
4.3.1 Challenges	43

Contents

4.3.2	Common variants	45
5	Synthetic data for deep learning	49
5.1	Data-centric AI	50
5.1.1	Common practices	52
5.2	Data collection	54
5.2.1	Discussion	55
5.3	Data generation	56
5.3.1	Computer graphics	57
5.3.2	Generative adversarial networks	59
5.3.3	Contributions	60
5.3.4	Discussion	65
5.4	Data augmentation	66
5.4.1	Image manipulations	67
5.4.2	Deep learning approaches	69
5.4.3	Contributions	71
5.4.4	Discussion	77
6	Conclusion	79
6.1	Contributions	80
6.1.1	Data generation	80
6.1.2	Data augmentation	81
6.2	Discussion	82
7	Outlook	85
	Bibliography	89
	Publications	117
	Paper A	119
	Paper B	147
	Paper C	163
	Paper D	169
	Paper E	193
	Paper F	201

1 Introduction

Datum is defined as anything “*that is given*”. Data, as all the facts and information that are given through observation or established as granted, have always been the driving force and the most essential element in the process of *learning*. Walking through history — from the earliest times of life and science to the modern artificial intelligence (AI) era — and the various forms of data, collection, storage tools, and analysis methods humans are still relying on data to distill knowledge.

Today it has become clear that the biggest challenge in the development and deployment of machine learning applications is the distribution of data that each algorithm is trained on. It is clear, that no matter the available computational resources and the potential of the deployed model, its success depends on whether the training data cover the distribution of the target domain. For example, an autonomous driving or robotic navigation system trained to detect vehicles of specific type or color, e.g., red cars, will fail to recognize a wider range of instances, like blue trucks or busses. Similarly, a face recognition algorithm trained on a limited range of skin color and facial characteristics is predestined to recognize particular ethnic, gender, and age groups excluding others, a behavior that under certain conditions can be misused and lead to discrimination incidents. Such failure cases of AI models, do not concern only the deficient variation in training data, but also the inherent bias that, most of the time, the training data carry. Thus, many machine learning applications learn from biased data, that additionally may correspond to just a part of the target domain, resulting in low fidelity, and potentially detrimental to the society, systems.

This thesis investigates how visual training data can be generated and used for machine learning applications. The foundation of the conducted research is described by one single idea: *synthetic image data* for supervised learning. For this purpose, image synthesis is approached from both a computer graphics and generative modeling perspective. It is furthermore examined as the sole source of data or as an augmentation strategy alternative. This chapter gives a brief introduction to data synthesis and deep learning. Then the thesis objectives are settled and finally, the book’s structure is outlined.

1.1 Visual data

Images are the product from a vision system, and constitute one of the most important ways to perceive the world. An example of such vision system is the human visual system (HVS). It is a complex structure that involves parts and functionalities of two major components: the eye, i.e., the sensor, and the brain, which is the computational and processing unit. The sense of sight is the ability to detect and process light. The physical scene, visible to the human eyes' visual field, is projected onto the retinas, where the light is converted to chemical substances and electrical impulses. The information leaving the retinas is finite and, in essence, corresponds to only a set of individual colored dots. The HVS processes the visual input before and after the incident light converts to colored dots, and as a result the generated mental images appear smooth and continuous.

The human need to freeze time, keep records, and eventually share visual information led to the invention of photography. A camera is an artificial vision system with many conceptual similarities to the HVS. In digital cameras, the incoming light focuses onto a silicon sensor, which resolution defines the captured image resolution, via the camera lenses. The light captured in the sensor's photosites is stored as electrical signals and then quantified as digital numeric values. After a series of processes, like demosaicing and noise reduction, the raw image is compressed and stored or used for various purposes [86, 193].

Captured images and sequences (videos) are the most popular forms of visual data, with other types to include views from multiple cameras and multi-dimensional data from laser scanners, or medical scanning devices. *Computer vision* algorithms aim to recover, analyze and understand such visual input, often as part of a decision-making process. Intuitively, computer vision attempts to automate tasks that the HVS can do. It has a long research and application history, and in the past decade, it has become one of the central fields in AI and machine learning. Among the reasons for this outburst is the democratization and wide accessibility of cameras — where high-resolution images can be nowadays captured with any conventional smartphone, and the consumer internet revolution. Typical computer vision tasks include face recognition, object detection, scene understanding for self-driving vehicles, and medical image analysis.

The importance of images and video in many applications has led to the development of numerous computer vision algorithms. Algorithms based on machine learning have shown to be both robust and generate accurate results. A central challenge, however, is that depending on the downstream task, the capturing process and the manual annotation effort and time can be expensive (often to the extent of infeasible [40, 51, 162, 272]) and prone to errors [152]. Image synthesis for training or testing data is, then, an appealing alternative, capable of providing large amounts of the required data type, along with a wide range of automatic and precise labels and absolute control on the generation and tuning process. Computer-generated images for supervised learning is a multidisciplinary task where computer graphics, computer vision, machine learning, and societal sciences intersect with several challenges. These range from data distribution definition

Figure 1.1: Generally, the constructing components of visual machine learning applications are *Data* and *Code*. Data includes all kinds of images that participate in the development. They can be captured (*real data*) and/or computer-generated instances (*synthetic data*). Code includes the algorithmic elements needed for the models to learn from the data. This thesis is a *data-centric* approach to visual machine learning application development, meaning that it focuses on the visual data and aims to improve the applications' performance by systematically improving them. Data-centric approaches are a relatively new concept and contrast to the, so far, conventional model-centric ones, where the development focuses on only improving the models to deal with the collected observations.

and diversity to introduced bias and artifacts inherent to the generation process. This thesis addresses some of these issues by exploring two image generation options for natural and histopathology images and presents a data-centric approach for visual AI applications. This means that the end machine learning applications improve their performance by systematically refining the underlying images used to train or evaluate a model than focusing solely on model development per se. Figure 1.1 illustrates the constructional components of visual machine learning and sketches this thesis domain.

1.2 Image synthesis

Similar to the HVS and digital cameras, image synthesis produces imagery defined by a set of colored dots (pixels). The colored pixels represent the virtual scene in various levels of realism and fidelity. This structure leads to an increased degree

of freedom when defining the data distribution for a machine learning application. Depending on the target requirements, the generated synthetic data can vary from low-level to photo-realistic representations. Moreover, the generation process provides control over the production line and the annotation needs. Digital image synthesis can rely on computational power, requires knowledge of physics, mathematics, and software development, and it does not simply end when an image file is created. It ends when a system “sees” the image, meaning that the generated imagery should conceptually and technically meet its characteristics, requirements, limitations, and restrictions.

Synthetic data for machine learning should always account for a twofold and interlaced domain gap. The first refers to the shift between computer-generated images and their real counterparts. They may need, for example, to comply with characteristics inherent to the capturing process, like noise or sensor-specific parameters. The other type is the deviation between the sets that a model learns from (training data) and is used at (testing data). This kind tightly relates with training data distribution coverage, a field where synthetic data can be a powerful asset for ordinary and long-tail cases. Consider a system that employs synthetic training images, alone or complementing captured instances, designed to detect dogs in photos. To be successful, the generated frames need to incorporate a range of optics-specific effects along with a variety of dog-specific features. The same applies if synthetic data serve as testing data, usually for system debugging and statistical analysis purposes. They need to simulate and cover, or even exaggerate, reality to reveal trends or weaknesses.

It is, therefore, clear that digital image synthesis has great potential to be a central tool in machine learning application development. This thesis approaches it from two perspectives. The *classical computer graphics pipeline*, and *generative image modeling*. Image synthesis with computer graphics has a long history with numerous advances that have enabled photo-realistic image generation [69, 185]. The conversion of data into images is a multi-level problem that can be divided into the two broad categories of *scene generation* and *rendering*. Scene generation is connected to the description and structuring of geometric and texture data on the computer, while rendering involves manipulating the data generated by the geometric modeling system to obtain an image that can be displayed on a graphics output device, i.e., a screen.

On the other hand, generative image modeling — specifically generative adversarial networks (GANs) [78] that this thesis deals with — is based on finding representations of given data distributions. GANs operate directly in pixel space and do not build the scene content or simulate physics for rendering as the computer graphics case. A learned generative model can generate and draw samples that are not in the training set but follow the same distribution. The rest of this section takes a closer look at each one of these image synthesis strategies. Chapters 3 and 4 will provide more detailed descriptions and technical details for both of them and showcase application examples.

1.2.1 Computer graphics

As mentioned, the pipeline for visual data generation can be divided into two main parts, *scene* or *content generation* and *rendering*. Scene generation refers to generating the features that build up the virtual environment where the sensors will be simulated. It describes all the two-dimensional (2D) or three-dimensional (3D) scene components and the data that need to be transformed into the output image. These data heavily relate to the geometric objects that represent or model physical objects. Geometric objects are most often represented by polygon (usually triangulated) meshes. The various geometries in a scene may be utilized and appear in different ways, depending on the content and the simulated sensor viewpoint. These transformations are performed by applying appropriate operations on the vertices of the meshes. The scene lighting is modeled through point or geometric light sources or illumination maps, while the simulated sensors can vary from a camera, lidar, or radar module to various scanner types. Scene generation extends to texture modeling and the general surface material definition and behavior, e.g., reflectance. This aspect plays a significant role in the final result and gets directly affected by the selected rendering procedure.

Rendering is the stage where all geometric, lighting, texture, and material models are combined and realized to the final synthetic image. It depends on the way cameras, or other sensors, are configured to measure and capture the virtual world and essentially simulates the light propagation in the defined virtual environment from the defined viewpoint. In short, each point on a surface in a virtual scene receives light directly from the visible light sources and indirectly from the reflected light from other objects. The latter indicates the recursive nature of inter-object reflection, as each object in the scene reflects a fixed portion of the incident light. This portion relies on the reflection characteristics of the material composition. This complex physics-based recursive algorithm is formulated by the radiative transfer equation or the rendering equation, in the absence of participating media like fog or fire, for which there are no analytical solutions apart from the trivial cases (see chapter 3 for details on light transport theory). Rendering is, therefore, carried out through light transport approximation methods that produce photo-realistic or non-photo-realistic results depending on their level of accuracy.

1.2.2 Generative image modeling

An alternative to digital image creation using classical computer graphics simulation is to use generative image modeling, which does not entirely depend on prior knowledge but uses data to learn their distribution and then draw / generate new samples from that same distribution.

In a broad definition, generative modeling refers to the statistical and probabilistic methods that take a collection of observations drawn from a distribution p_{data} , i.e., a training set, and learn to represent an estimate of this with a probability distribution p_{model} . This can be done either *explicitly*, meaning that the model learns to perform density estimation and the learned probability distribution p_{model} can be evaluated for a

particular value x to obtain an estimate $p_{model}(x)$ of the true density $p_{data}(x)$, or *implicitly* where the model is only able to generate samples from the estimated model distribution p_{model} .

Explicit models are usually likelihood-based and learn directly the distributions probability density function via (approximate) maximum likelihood. Typical likelihood-based examples are autoregressive models [76, 138, 258], normalizing flow models [58, 59], energy-based models [141, 228], and variational auto-encoders (VAEs) [131]. In implicit models, the probability distribution is represented by a model of its sampling process [171]. The most prominent example is GANs, where new samples from the data distribution are synthesized by transforming a random Gaussian vector with a neural network.

1.3 Deep learning

The desire for “*thinking*” machines is not new in the history of this world. Centuries before technology made self-moving devices possible, the tales of the first on earth giant bronze robot Talos, artificial woman Pandora, and their creator god, Hephaestus, filled the imaginations of people in ancient Greece. Greek mythology was exploring ideas about creating artificial life and grappling with still-unresolved ethical concerns about bio-techne (bios = life, techne = crafted through the art of science). Moreover, Roman, Indian, and Chinese myths envisioned artificial life, automata, self-moving devices, and human enhancements [163]. As Adrienne Mayor says, “*There is a timeless link between imagination and science*”. Fast forward to 2700 years later, and in a nonfictional framework, AI is nowadays better defined, continuously developing, and most importantly, realized.

The set of concepts and algorithms that we today call *deep learning*, has a long and rich history dating back to the 1940s. From its origin till today, the relevant progress and advances were not always known and grouped under this name. There are three distinct development waves, with the first one to be *cybernetics* (~1940–1960) related to theories of biological learning [165] and implementations of the first models, such as the perceptron [204], enabling the training of a single neuron. The second wave started with the *connectionist* approach or *parallel distribution processing* (~1980–1995), when back-propagation [205] enabled training a neural network with one or two hidden layers. The current and third wave, identified as deep learning, started in early 2000s [24, 98, 99] when training deeper neural networks became possible. Today deep learning research is one of the most active research areas, with tremendous progress and a high-speed rate of development. Algorithmic, training data and computational power advances have brought deep learning solutions to the spotlight for various cognitive and perception applications. Despite the short time of this wave, deep learning research has already showcased impressive progress to both unsupervised and supervised learning. Finally, it is worth mentioning the connection of deep learning with neuroscience within this chronology. The earliest forms of deep learning were simple linear models motivated from a neuroscientific perspective. However, modern deep learning goes beyond that,

and neuroscience, even though has been a source of inspiration is no longer considered its driving force. It is more accurate to think of deep learning methods and models as approximation tools that try to achieve statistical generalization rather than models of neural brain functionalities.

Deep learning is an approach to AI and a part of the general machine learning realm [79, 142]. It is defined to allow computers to learn from experience, i.e., observations or else *training data*, and understand this input through a hierarchy of concepts. In this nested, *deep* hierarchy each concept is defined through its relation to simpler, lower-level ones, and therefore the computer can *learn* complicated concepts by building them out of simpler blocks. For example, an image of a dog can be broken down into simpler components, like the animal's outline, defined in turn by edges.

This thesis focuses on supervised learning, where the algorithms require labeled visual data relevant to the task and application. In supervised learning, each image must be accompanied by a ground truth annotation that complies with the learning task at the time. Once more, consider the dog detection example. A deep learning solution for this would require images where each instance would come with a bounding box. Other conceptually similar recognition applications are image classification where each image is followed by a name label that indicates if any relevant class is present, and semantic segmentation that masks the imagery in semantic groups and requires pixel-level class annotations. The spectrum of image recognition applications spans from everyday life scenarios to medical imaging tools and maintenance, repair, and operation industry. Deep learning approaches are widely embraced in this domain, showcasing numerous commercial and non-commercial solutions. Other active application areas are advertising, financial services, language processing, recommendation engines, and many more.

The third wave and rise of deep learning evolved with the wide accessibility to computational resources, which enabled larger models, and the expanding of training data sizes. The effort required to make the most of a deep learning algorithm decreases as the amount of training data increases. Modern deep models often use the same core learning algorithms from the 1980s with the difference that they feed them with higher quality and more training data, a direct benefit of the digitization age. Visual machine learning is now part of everyday life, as it is the base or component of a great gamut of solutions, services, and decision-making systems. Image data readiness and curation is a central aspect of deep learning research and application development.

1.3.1 Training data

Computational power can limit deep learning models in terms of actual capacity and training time. Apart from this, there are two components for successful application development. The first one is determining the network architecture, meaning the overall network structure and hyper-parameters. This step includes defining the number of layers and their type, the units per layer, initialization of weights, loss function and optimizer selection, learning rate schemes, regularization terms, and more. The other

one is the *data* used for training the defined network for the defined downstream task. Training data are the observations from which the network will learn patterns and trends to generalize to the application domain. It is obvious that a model will not perform as expected if the training data cannot represent appropriately the distribution of data that is supposed to be covered.

Training deep networks always involve finding a balance between these two factors. A model with too little capacity cannot learn the problem, whereas a model with too much capacity it is possible to learn too well and overfit the training set. On the other hand, too little or low quality or out-of-the-target application training data will lead to poorly performing models no matter the capacity and the data regularization methods. This thesis emphasizes the training data factor in visual deep learning advancement and proposes image synthesis as a data resource. Synthetic training data can enhance or replace captured data, if needed, and act as a balancing term for the domain gap between the source and target domains. The control that image synthesis provides over the generation process is essentially control over the training data. Apart from the ability to define the content and characteristics, there are enabled multiple image and annotation modalities at minor or no human effort cost. Even though this thesis steers attention to synthetic training data, the potential of synthetic testing data should not be left unmentioned. Computer-generated images of high quality can be an attractive alternative for thorough statistical analysis and benchmarking.

1.4 Objectives

This thesis contributes to the data component of visual machine learning, driven from both research and applicability, and aims to provide the learning algorithms with the resources to succeed, employing synthetic image data. The included projects engage different image synthesis methods and cover two distinct image modalities to explore the span of the potential of synthetic data. Moreover, the generated data are evaluated in two training modes. Figure 1.2 summarizes and interconnects the contribution areas. The relevant papers are listed and briefly described, along with the author's contributions to each one of them, at the beginning of this book (pages vii and ix respectively). In this section, they are grouped under the three aforementioned research areas.

Image synthesis methods

The papers of the thesis study and present synthetic imagery using *classical computer graphics* (Papers **A** and **B**) and *generative image modeling* (Papers **C**, **D**, **E** and **F**).

The author was first introduced to the idea of synthetic images for deep learning applications with the rise in the development of autonomous driving back in 2016 and the collaboration with 7D Labs Inc. that led to Paper **B**. This introduces a novel approach for generating images for urban scene understanding. It is based on procedural modeling and physically-based rendering and produces a new

Figure 1.2: The thesis objectives are threefold and relevant to the data domain for visual machine learning. The central idea is synthetic data, that act either as the sole source of training data in deep learning models (papers **B** and **F**) or as a finetuning or augmentation strategy, mainly in cases of training data scarcity and imbalance (papers **B**, **C**, **D** and **F**). Image synthesis is explored via classical computer graphics methods and generative modeling, GANs in particular, while the targeted applications are recognition tasks related to real-life scenarios and digital pathology.

virtual world per frame. The pipeline provides high levels of flexibility and scalability while creating highly realistic, content-rich results. Consequently, Paper A presents a thorough literature review of all other computer graphics methods that had produced training data for visual machine learning. It presents a taxonomy based on the used modeling and rendering techniques and a computer vision task categorization. Moreover, a relative quality index accompanies each method/dataset that reflects reported performance combined with data complexity measures and aims to serve as a learning potential hint.

Papers C, D, E and F approach image synthesis via generative modeling. Papers C, D and E create data complementary to captured training sets, i.e., augmentation data. In these, the final image generation is achieved through image transformation from one source domain to the target application domain, employing image-to-image translation and CycleGANs [294]. Papers C and D tackle the problem of metastasis detection of colon cancer in the lymphatic system when very little or no target domain training data are available. They suggest using colon primary tumor and already available annotated tissue from cancers in different places (breast and skin) that exhibit contrasting levels of similarity to the target domain tissue. These out-of-target domain training data are evaluated as augmentations both in their original/raw form and as the source domain for the domain adaptation with CycleGANs. The results show not only improved performance, but also increased models' robustness. Paper E investigates GAN variants synthetic training image augmentations for multi-class weather classification from a single image when there are available very little and class-imbalanced data. Here, the underrepresented class is complemented either with intra-class classical augmentations, or intra-class generated data using a noise-to-image GAN, or inter-class generated data using the vanilla and a modified CycleGAN. The latter proves to be the most efficient generation method for the task. Finally, Paper F suggests ensembles of independently trained GANs — two variants and various ensemble sizes — to generate synthetic data of increased diversity that are capable of training deep classifiers in isolation from real data.

Image modalities

The image synthesis investigation for machine learning conducted in this thesis aims to cover two basic image modalities: *natural images* (Papers A, B, E and F) and *histopathology images* used in digital pathology (Papers C, D and F).

Natural images are content-rich, high variance representations of objects, instances, and events. They can be roughly and intuitively defined as examples that could be captured in reality, e.g., nature, urban or domestic scenes. Under this thesis context, the notion of synthetic natural images may sound like an oxymoron, but the interpretation is computer-generated representations of real-life or imaginary concepts that are characterized by extensive diversity and structural variance. Papers B and E synthesize natural images for urban scene understanding and weather classification respectively. Each one of them addresses the generation process differently to achieve high levels of outcome heterogeneity.

In Paper **B** the synthetic natural images are used as the sole source of training data and for captured data fine-tuning, while in Paper **E** the created weather-depicting imagery work in combination with captured images and act as complements to underrepresented classes. In addition, Paper **A** gives a comprehensive, for the time being, list of datasets and methods that produce mostly natural images for training computer vision applications. Paper **F** showcases low resolution and complexity natural images generation utilizing ensembles of GANs and points out that independently trained GANs are seemingly on par with boosting methods while generating diverse and not overfitted results.

Digital pathology is an image-based pathology field that acquires, manages, and interprets information from digitized glass slides of tissue. This imaging modality is the base of deep learning-based diagnostic and decision support tools for pathologists, especially for various types of cancer. Due to the cellular structure, this modality exhibits a seemingly lower extent of diversity compared to natural images, and the acquired data may seem similar even for different conditions. Apart from tissue resemblance, this is dependent on the acquisition and scanning process that can also introduce some heterogeneity but in a relatively controlled way. Papers **C** and **D** generate pathology data for colon adenocarcinoma metastasis detection in the lymph nodes. The works were done in sequence and utilized intra-organ data, i.e., colon primary tumor, and inter-organ tissue, i.e., tissue data from breast and skin, to synthesize image samples that correspond to the target domain distribution (lymph node colon cancer). The annotation-effort-free synthetic data augment limited training data scenarios and prove to be an efficient strategy, boosting both performance and robustness. Paper **F** generates histopathology patch images for breast cancer detection. This is done in an anonymization of a real set fashion, utilizing ensembles of independently trained GANs.

Training modes

The generated data are evaluated both in conjunction *with* (Papers **B**, **C**, **D** and **E**) and *without* real/captured training data (Papers **B** and **F**).

Papers **C**, **D** and **E** handle synthetic data as a form of complement (augmentation) to captured training sets. These papers deal with low-data scenarios, where the captured data participate in the generation process of the synthesized imagery through image-to-image translation. Then, the created images act as an augmentation set that expands the training set's diversity and size. The real data are leveraged to their maximum potential and complemented by new, synthetic samples drawn by the source data distribution. This enables usability and applicability of even small training sets, where the set size and the distribution coverage are extended. Paper **B** presents an image synthesis method and evaluation where the generated images are used both as the only training data source and as a means of fine-tuning on real data. Training only on purely synthetic data outperformed the other, at the time, image synthesis methods but did not perform as well as training on real images. Fine-tuning on captured urban scenes proved to be the most effective approach. Paper **F** also assesses synthetic images as the

only source of training data. Here, it is amplified that image synthesis could be an alternative when privacy and ethical issues arise, for example, for sensitive medical visual data for anonymization purposes. Similarly, paper A presents a thorough study on synthetic datasets and image generation methods that act as a training data source in various computer vision tasks.

1.5 Outline

This chapter introduced the fundamental concepts of this thesis and motivated synthetic data and their benefits for data-centric visual deep learning. It outlined the two image synthesis approaches that this work focuses on and some basic concepts around deep learning and training data for supervised tasks. Finally, it established the objectives of the thesis and contextualized the contributions.

What follows is conceptually divided into two parts. The first part consists of the following chapters that will give more details on the necessary background on deep learning operations, computer graphics, generative modeling and discuss visual data generation in connection with the papers included in this thesis. The second part consists of the papers themselves, published throughout the time of the presented doctoral journey.

Chapter 2 provides general background on deep learning, neural networks, and basic learning concepts, necessary to understand the framework of the conducted experiments and the significance of data in the process. It also defines visual machine learning for computer vision and gives a notion of the rich domain of applications that use synthetic training data to solve visual tasks. This part complements paper A, which is the complete thorough literature review of image synthesis methods for visual machine learning. Finally, digital pathology is introduced to familiarize the reader with this application domain.

Chapters 3 and 4 constitute the building blocks of image synthesis methods described in this work. Chapter 3 focuses on computer graphics, providing details on both modeling and rendering techniques, while chapter 4 provides all the necessary framework around generative image modeling, with a special emphasis on GANs and their variants used in the thesis.

Chapter 5 connects the dots into a concrete whole around synthetic data for visual deep learning. The desired data qualities for a data-centric pipeline are defined, and the benefits and potential of synthetic data over them are highlighted. Then, the two contribution areas are extensively discussed. Section 5.3 elaborates on training data generation for supervised deep learning applications using both computer graphics and GANs and incorporates the presented approaches for diversity and privacy in natural and histopathology images, respectively (papers B and F). Section 5.4 develops around data augmentation and associates with it the included in this thesis work on augmentations for data scarcity in digital pathology (papers C, and D) and class balancing in

natural images (paper E).

Chapter 6 gives a unified summary of the presented work, and thesis contributions collectively, surrounding them with a discussion around the related fields and general concepts. Finally, chapter 7 ends this first part with an outlook to the potential future directions of synthetic data as a pivotal part of data-centric visual machine learning.

2

Background

Having set the motivation, framework, and objectives of the thesis, it is essential to provide the necessary theoretical background that supports the parts of the conducted research both in the conceptual and in the more technical perspective.

Over the past years, machine learning, including deep learning, is steadily one of the most animated research fields, producing several advancements in various application fields. During these long-standing efforts, it has been proven that deep learning pipelines are capable to undertake complex learning tasks. Deep learning algorithms are built from interconnected groups of nodes, formulating networks, that allow for a learning process of gradual complexity. To ensure their successful development and deployment, several factors, concepts and strategies has been investigated and tailored, always with the purpose that the underlying learned model shall perform well when encountering unseen data.

Most relevant under the current context of synthesizing visual data for machine learning are learning algorithms that solve tasks related to visual perception, i.e., computer vision tasks. Computer vision is the field that enables computers to extract information from images or videos and interpret the depicted “world”. Computer vision powered by machine learning (*visual machine learning*) is an active research area for several tasks, like identifying and classifying objects, showing great potential for relevant decision-making systems. For example, digital pathology with deep learning, which aims to “understand” digitized pathology data, is a rising application field that has already shown promising results and even got approval for actual clinical use [1]. It is therefore obvious that synthesizing images for supervised visual machine learning is an open field that can accommodate various image modalities and provide annotated data for a wide range of applications.

This chapter discusses the various aspects of a deep learning pipeline designed to undertake computer vision tasks. To this end, a description of neural networks is provided first. Later on, the basic concepts such as dataset partitioning, imbalanced datasets, generalization, underfitting and overfitting, model and dataset bias, domain shift, and regularization are discussed. Finally, a review of computer vision areas that benefit from synthetic training data, along with a discussion for digital pathology, in particular, is provided.

2.1 Deep learning

AI is witnessing tremendous growth in bridging the gap between the capabilities of humans and machines. In the long quest of evolving and establishing it as a mature tool, deep learning models have proven their ability to solve tasks that are particularly difficult to describe rigorously and formally. These tasks are often easy to perform by a human operator, for example, classify images based on their content. However, their automation through deep learning pipelines is required to be deployed on a large scale and support further transition to fully digitalized services and products.

Deep learning algorithms are formulated as interconnected groups of nodes — *neural networks* — that allow learning out of blocks of increasing complexity. This section gives a closer look at the fundamental neural networks types, basic learning concepts, and deep learning applications.

2.1.1 Neural networks

The most basic deep learning model is the **deep feedforward neural network**, or multilayer perceptron (MLP). It is a mathematical function composed of a series of simpler ones structured in layers, where each layer provides a new representation of the input. This way, the initial input is mapped to output values. For example, a classification problem requires a function f^* to map an input sample x to a category y , i.e., $y = f^*(x)$. A feedforward network defines a mapping $y = f(x; \theta)$ and learns the set of the parameters θ that results in the best function approximation. In MLP, nonlinear functions of the input can be represented by extending linear models. This is done by applying the linear model not directly to the input but to a transformed input $\psi(x)$, where ψ is a nonlinear transformation. The model is then expressed as $y = f(x; \theta, w) = \psi(x; \theta)^\top w$ and it is therefore possible to learn ψ through the set of parameters θ from a broad class of functions. Then the learned parameters w linearly map $\psi(x)$ to the output values. In this setup ψ defines one hidden layer of the deep feedforward network. The word *feedforward* indicates the information flow in these models; from the input x , through the intermediate computations ψ and f to the output y .

Convolutional neural networks (ConvNets or CNNs) are another fundamental class of deep neural networks [139, 140]. They are specially used for grid-form data, like images and time-series, and have showcased impressive results in computer vision applications [91, 135, 227, 235]. As their name implies, this type of network utilizes convolution, most often applied to many layers, to capture the spatial and temporal dependencies in the input data through the application of relevant filters. CNNs building blocks are convolutional and pooling layers in the first stages of the network and fully connected layers in its final part. The convolutional layer is where the majority of computations happen. A feature detector, also known as a kernel or a filter, is a set of weights that moves across the receptive fields of the layer input, checking if the feature is present by operating convolutions. The output of these local, weighted sums is called feature maps. These are then passed through a nonlinear transformation, such

as the Rectified Linear Unit (ReLU) function. Similar to the convolutional layer, the pooling operation moves a filter across the input. However, this filter does not have any weights but applies an aggregation function to the values. A pooling operation computes the maximum, or the average, of a local patch of units in one or more feature maps. This operation reduces the dimension of the representation while aiming to merge semantically similar features into one. The information loss is weighed out by the reduced complexity and the increased invariance to small shifts and distortions (leading to lower overfitting risk). Deep CNNs stack convolution, nonlinearity, and pooling layers followed by fully connected layers. Thereby, the network builds a hierarchy where the higher-level features are obtained by composing lower-level ones. In images, for example, the later layers can see pixels within the receptive fields of prior layers, and the lower-level patterns can construct object parts, which in turn assemble to higher-level full object descriptions.

Due to their potential and performance, several CNN architectures have been developed and tested in various kinds of datasets. Widely used networks include architectures like AlexNet [135], VGG-16 [227], GoogleNet [235] and ResNet [91]. Khan et al. [127] provides a detailed survey of the existing CNNs and their evolution. This thesis employs simple forms of this type of network to showcase the applicability of the studied data generation methods.

Extended feedforward neural networks that include feedback connections are called **recurrent neural networks** (RNNs), and they are the third central family of deep learning models [205, 206]. They are a type of network that uses, similarly to CNNs, sequential data or time-series data. Their main difference is a “*memory*” feature that they account for, as they allow for previous outputs to be used as inputs while having hidden states. In other words, the output of recurrent neural networks depends on earlier elements within the sequence. In this category, there are three basic design choices. The first is a network that produces an output at each time step and draws recurrent connections between hidden units. An alternative design creates a network with output at each time step but recurrent connections only from the output of one time to the hidden units at the next time step. The third pattern is a network with recurrent connections between the hidden units that read an entire sequence and then produce a single output. In all cases, RNNs share parameters across each layer of the network. In contrast with feedforward networks, where there are different weights across each node, recurrent neural networks share the same weight parameter throughout all layers. When they are unfolded in time, RNNs are deep feedforward networks in which all the layers share the same weights.

Neural network learning is essentially an optimization process. The network learns to map a set of inputs to outputs. Learning the weights of a neural network is a search or optimization problem, where an algorithm navigates the space of possible weight values that the model can use to make accurate predictions according to the task. Typically, a neural network is trained using the stochastic gradient descent (SGD) optimization algorithm. The term gradient refers to error gradients. During training, the model uses

the set of weights at the current iteration to make predictions. Then, the error between these and the ground truth is calculated. SGD seeks to change/update the weights so that in the next evaluation the error is reduced. This way, the optimization algorithm navigates down the gradient (or slope) of error.

The error, and thus a candidate solution for the weights, is evaluated through a function called objective or cost or loss function. This needs to be selected/designed to concentrate all the aspects of a possibly complex system down to a single scalar value used to rank and compare possible weight solutions. Determining the correct loss function for each task can be challenging as it needs to capture the properties of the problem and be motivated by concerns that are important to the project and stakeholders. Popular choices have been the mean square error loss, binary cross-entropy for binary classification, and multi-class cross-entropy for multi-label classification.

The nonlinearity in deep neural networks affects various types of loss functions to become nonconvex. This further results in the training of networks by iterative, gradient-based optimizers. *Forward propagation* indicates the information flow in the network: the initial input propagates to the hidden layers and finally produces the output. During training, it continues until the scalar cost is computed. *Back-propagation* allows the cost information to flow backward to compute the gradient [205]. It is based on the chain rule of calculus that performs operations in a specific order to improve computational efficiency. Finally, the optimization algorithm, e.g., SGD or Adam [130], uses the calculated gradient in the learning process. Back-propagating gradients in CNNs is as simple as through a regular deep network, allowing all the weights in all the filters to be trained. Recurrent neural networks use back-propagation through time (BPTT) to determine the gradients, an algorithm that is specific to sequence data. The principles of BPTT are the same as traditional back-propagation, where the model trains itself by calculating errors from its output layer to the input layer. These calculations allow to adjust and fit the parameters of the model appropriately. BPTT differs from the traditional approach in aggregating errors at each time step, whereas feedforward networks do not need to sum errors as they do not share parameters across each layer.

2.1.2 Basic concepts

Building successful deep learning pipelines is a non-trivial task associated with a handful of concepts and respective challenges. In the current section, basic learning-related concepts, such as the composition of visual data sets, domain shift, regularization, and generalization, are briefly discussed.

Dataset partitioning

End-to-end training and evaluation of a model require a training and testing set. It is essential for these sets that they do not overlap and consist of the same images. The training set should include target distribution samples but not identical to the testing set. A common practice for this is to split the available dataset into subsets. *Dataset*

partitioning is a simple principle where the provided data are divided into three sets. The first is the training set, which is used to fit the model and adjust the parameters. The second one is the validation or development set, which consists of images that are held back from the training set. This subset provides an unbiased evaluation of a model fit on the training set while tuning model hyperparameters. It is used for frequent evaluation (over the training iterations), where the model occasionally sees the data but never learns from them. However, its results can indirectly affect the model as they are used for hyperparameters updates. Finally, the third is the test set that is the data that is employed to provide an unbiased evaluation of the final model. This one is used once the model is completely trained (using the training and validation sets).

Setting up the training, validation, and test sets can highly impact model development. Each of them, and thus the original dataset, too, should be suitable representations of the domain of the learning task. It is important to choose the validation and test sets from the same distribution, and it must be taken randomly from all the data. These sets need to reflect the data that the model expects to get in the future. Their size should be “sufficiently large” for their results to be representative of the model performance. An example of a typical proportion of these partitions is 70/20/10 (training/validation/testing).

Imbalanced data

A relatively common problem in supervised learning, which requires annotated data, is the lack of availability of data representing specific categories of a downstream task. Even with “sufficiently large” datasets, data scarcity can occur in specific classes of the original set, leading to *class imbalances*. In practice and considering image classification problems, class imbalance means that some specific output categories are associated with a small amount of visual data, whereas abundant data exist for other output categories. As a result, incomplete representations exist for specific parts of the learning tasks domain, i.e., for the ones related to the specific output categories, and hence the overall training process does not lead to trustworthy results. Image synthesis poses a straightforward solution to this problem, producing on-demand visual data for the specific output categories associated with scarce data.

Generalization

The principal challenge in deep learning, as well as in general machine learning, is to construct a model — trained on a sufficiently large, yet finite, set — that performs well on other unseen datasets. This ability to perform well on previously unobserved inputs is called *generalization*, while the associated error metric is called generalization error.

Considering the given training set, an error metric is introduced, namely, the training error to monitor the learning process of the model on the training data. As expected, the minimization of the training error is the objective of a defined iterative process within the training phase. However, this minimization is not enough to guarantee the good

performance of the model and considerations shall be made to maintain the generalization error low as well. The latter is a multi-factor challenge, associated with most of the basic concepts and methods related to the development of data-centric, deep learning pipelines, such as underfitting, overfitting, transfer learning, and regularization.

Underfitting and overfitting

Following the standard machine learning pipeline, three main steps are extracted: 1) sample the training set, 2) define model parameters, and 3) sample the test set. The successful implementation of these steps will lead to the minimization of the training error and the associated gap between the training and testing error. However, success is associated with overcoming two main challenges: *underfitting* and *overfitting*. The first occurs when a learning model is not able to achieve a low training error, whereas the latter occurs when the test error is significantly larger than the obtained training error. Essentially, these two challenges necessitate selecting the proper capacity of the model, i.e., the ability to fit a wide variety of functions, always based on the underlying complexity of the learning task. Further to that, models with low capacity will struggle to properly fit the training dataset, while models with high capacity will map the properties of the training dataset that are not beneficial when generalization is requested in the test dataset.

Model and dataset bias

Underfitting and overfitting can be described from the *bias and variance* of a deep learning model and their well-known tradeoff. In essence, bias and variance express two different types of errors of the model. Bias error measures the expected deviation from the true value of the function or parameter, while variance measures the deviation from the expected model value that any specific sampling of the data is likely to cause. Careful selection of the model capacity, including the definition of deep learning model parameters, leads to successful negotiation of the bias-variance tradeoff, where both bias and variance errors remain low. Moving away from the optimal tradeoff, high variance error implies high sensitivity of the model in respect to small fluctuations of the training dataset that are not beneficial when generalizing, i.e., overfitting. On the other hand, high bias error means that the expected deviation of the model, considering the ground truth data, is large, which is a direct indication of the inability of the model to capture the dynamics of the learning task.

On top of the model bias, as described above, *dataset bias* is also an additional issue that arises when data-centric deep learning pipelines are deployed. In the case of data bias, the content of the training set is the problem itself. In short, the data bias means that the training dataset contains visual information that strongly pertains to a specific class or category of interest. Thus, the biased dataset does not accurately represent the different types of visual data that are expected to be the subject of the requested learning task in the deployment stage. As expected, the data bias will lead to poor generalization of the

deep learning models. A glimpse of superior performance can be expected in case some test sets contain a similar class of visual data compared to the biased training dataset.

Domain shift

A challenge of discriminative, predictive, and generative models is to achieve good generalization under *distributional shifts*, i.e., changes in the distribution of the data seen in the deployment stage compared to the distribution of the training data. Such shifts, namely domain shifts, can deteriorate generalization performance, even though the underlying learning task remains the same and the training dataset is large.

There are two main methods for improving the generalization performance under domain shifts: transfer learning and domain adaptation. In the first, the model is pre-trained on a related task with a large dataset and then finetuned on the new, shifted data. This two-step approach can reduce the risk of overfitting to the source domain because convolutional features learned on larger datasets are more general. The second method, i.e., domain adaptation, focuses on adapting the model after the training phase to improve its accuracy on a specific target domain. The latter is achieved by minimizing the distance between the source and target feature distributions in some fashion.

Regularization

Collectively, *regularization* refers to any modification made to a learning algorithm that aims to reduce its generalization error but not its training error. In the context of deep learning, the majority of the regularization strategies are based on regularizing estimators. Regularization of an estimator works by trading increased bias for reduced variance, thus attempting to reach closer to the optimal value of the bias-variance tradeoff. In that fashion, an effective regularization strategy makes a profitable trade, reducing variance significantly while not overly increasing the bias.

2.1.3 Applications

Deep learning models are of extreme importance to machine learning, and they are the basis of many commercial and non-commercial applications. There are numerous examples, including financial services and fraud detection, political trends with demographic and election predictions, recommendation engines and hyper-targeted advertising, self-driving vehicles, various kinds of robotic systems, virtual assistants, automatic text generation, and many more. Referring to the network families described earlier, recurrent neural networks are commonly used for ordinal temporal problems, like natural language processing (NLP) and image captioning, while deep convolutional neural networks specialize in classification problems and all kinds of computer vision tasks, e.g., non-destructive testing and various types of quality inspection for automated manufacturing processes and predictive maintenance pipelines.

Taking a closer look at visual applications that this thesis is closely related to, the spectrum of the fields using deep learning solutions is immense. Basic tasks like

object detection, human body or face recognition, semantic segmentation and scene understanding, image and object classification, with or without localization, and image colorization, stylization, reconstruction, and super-resolution are nowadays typically solved with CNN approaches, including GAN variants. They are widely used in manufacturing, agriculture, entertainment, virtual reality, transportation, and healthcare. With this wide range of utilization areas, it is clear that the requirements on training, and testing, image data of diverse content, quality, and formats are massive. This thesis presents synthetic image generation as a data source and exemplifies its applicability in autonomous driving, cancer detection, and weather classification.

2.2 Computer vision

Visual machine learning relates to tasks of visual perception and includes algorithms that utilize visual training data. To a great extent, in practice, visual machine learning solves computer vision tasks, from low-level ones, like visual odometry and optical flow, to higher-level tasks, like 3D scene layout, semantic understanding, and object detection and tracking.

Paper A presents a thorough investigation of the image synthesis methods and possible associated released synthetic datasets that have been developed and employed with the emergence of deep learning. For these methods, the greater growth came in connection with the rise of mainly recognition applications over the last decade. Paper A provides two types of taxonomies. One is related to the data generation methods with a focus on the computer graphics perspective, and the other is relevant to the computer vision application fields. Moreover, it defines qualitative criteria to rank the various synthesis techniques and synthetic sets within the different computer vision application areas, focusing on data complexity and performance aspects.

To further highlight the usefulness of image synthesis for deep learning and lay the groundwork, in this context, for the type of applications that use them, this section gives an overview of the active computer vision areas that make use of and benefit from image synthesis methods as a source of generating training data. Following paper A, these computer vision areas are summarized to:

Feature-based alignment that estimates the motion of 2D or 3D feature correspondences between images. Human body and object pose estimation are typical examples that have utilized synthetic training data.

Dense motion estimation that determines the motion between subsequent frames. Optical flow, due to the labor-intensive annotation process, has extensively used synthetic data in machine learning solutions.

Stereo correspondence estimates 3D models of a scene from 2D images. Relevant tasks that benefit from training data generation are depth, disparity, and scene flow estimation.

Recognition that is one of the larger computer vision areas with extensive research

on scene understanding. This can relate to identifying contextual components and detecting specific objects/features. Common applications that employ image synthesis for training data include semantic, instance, and point-cloud segmentation, object detection, and class and face recognition.

Structure from motion that estimates locations, structures, and egomotion from image sequences. In this category, synthetic training sets have been utilized in visual odometry.

Computational photography and image formation that use image analysis and processing to create images that go beyond the capabilities of traditional imaging systems and study the constituting elements of an image, respectively. Within this framework, camera design, noise modeling, and intrinsic image decomposition have utilized synthetic data.

Table 2.1 shows the computer vision areas and applications categorization that utilize synthetic training data presented in paper A. Here is reformulated to indicate the various image synthesis methods according to the data generation method discussed in this thesis, i.e., computer graphics or generative modeling. Moreover is extended to include a few methods that employ both computer graphics and generative modeling techniques to generate visual data for training machine learning solutions (*hybrid techniques*). Importantly, the formulation is restricted to techniques that intend to create entire training sets and therefore does not include methods specific to, e.g., data augmentation or synthetic image generation only for testing purposes.

It is worth mentioning that the presented methods are listed under the tasks that are reported in their original publication. However, it is possible that several of the included data generation methods and synthetic datasets can be applied to solve other tasks too. That is, one type of ground truth for supervised learning, suitable for a certain problem, can often be used to produce data for other applications. For example, semantic segmentation labels can be used to provide object detection bounding boxes.

Finally, this categorization should not be perceived as a static, limited portrayal but rather as a dynamic list where the balance of applications is subject to change as new fields and example methods are added. The goal of this taxonomy is to emphasize that visual training data generation is an evolving field. More synthesis methods and their visual outcomes will keep being added and extend the presented computer vision areas and applications, all with the same common goal to boost the performance of the learned models providing wider feature coverage and overall generalization.

2.3 Digital pathology

Digital pathology is about acquiring, managing, sharing, and interpreting pathology information in a digital environment. This formulation has the potential to improve patient care and support the pathology workforce by making the diagnosis and monitoring of disease much more efficient. Digital pathology visual tasks with deep learning, e.g., cancer detection, is in essence typical computer vision tasks applied to this specific type of image instead of natural images. This is an emerging direction that has already

		Computer graphics	Generative modeling
Feature-based alignment	Pose estimation	[232], [238], [245] ¹⁷	
	Human body	[223], [187] ¹ , [182], [44] [259] ⁸ , [296], [65] ⁹ , [241]	[224]
Motion estimation	Optical flow	[19], [41] ¹³ , [87] ² , [60], [161] ⁵ [71] ⁴ , [109], [201] ⁷ , [259] ⁸	
Stereo correspondence	Disparity estimation	[41] ¹³ , [161] ⁵ , [291]	
	Depth estimation	[41] ¹³ , [87] ² , [10] ³ , [71] ⁴ [259] ⁸	
	Scene flow estimation	[166], [161] ⁵	
Recognition	Semantic segmentation	[87] ² , [10] ³ , [203], [71] ⁴ , [200] [90], [89], [164] ¹⁴ , [290], [95] ⁶ [201] ⁷ , [259] ⁸ , [12], [274] ¹² , [20] ¹⁶ [122] ¹⁰ , [128], [107] ¹¹	[52], [221], [122] ¹⁰ , [202]
	Instance segmentation	[71] ⁴ , [164] ¹⁴ , [95] ⁶ , [201] ⁷ [274] ¹² , [107] ¹¹	[197]
	Object detection	[159], [187] ¹ , [71] ⁴ , [119], [164] ¹⁴ [190], [245] ¹⁷ , [244], [274] ¹² , [189] [122] ¹⁰ , [107] ¹¹ , [100], [11] ¹⁵	[122] ¹⁰
	Object tracking	[71] ⁴ , [65] ⁹ , [201] ⁷ , [11] ¹⁵	
	Classification	[122] ¹⁰	[246], [122] ¹⁰ , [104]
	Point-cloud segmentation	[285]	
	Face recognition	[270], [2], [133]	[247], [74]
Structure from motion	Visual odometry	[201] ⁷	
Computational photography & Image formation	Camera design	[33], [154], [153]	
	Noise modeling	[145], [3], [28]	
	Intrinsic decomposition	[23], [196], [148], [20] ¹⁶ [21], [266], [27], [225]	

Table 2.1: Computer vision applications that benefit from synthetic training data, along with associated image synthesis approaches (superscripts denote cross-application approach).

shown promising results. This section provides basic informative background on the field as an application area of this thesis.

A core part of pathology is the ability to interpret images of tissues and cells at a higher resolution than the naked eye. For a very long time, the microscope was the only available means to this end. Digital cameras and microscope-mounted video cameras led to telepathology, while whole glass slide scanners enabled, just the past two

decades approximately, what we call today digital pathology. To give some context, the surgically removed tissue first undergoes chemical treatment, sectioning, and staining, most commonly with hematoxylin and eosin (H&E) to highlight important features and enhance the contrast, and then is placed over glass slides and scanned at different magnifications, e.g., $\times 20$, or $\times 40$. Naturally, the tissue preparation and scanning processes can introduce artifacts and biases. The resulting images are called whole slide images (WSI), and they are in the order of gigapixels, i.e., tens or few hundreds of thousands pixels per dimension, requiring some gigabytes of memory each.

WSI are digital files of constant quality, in contrast with physical tissue specimens, that are examined by pathologists using specifically designed image viewer solutions [7, 67]. These are usually capable of further magnification options while providing at the same time several tools and conveniences for display views, navigation, orientation, and annotation to support a fast and accurate diagnostic process. WSIs are an immediate benefit of the digitization era that can be made available instantaneously to multiple examiners in different parts of the world without the need for a microscope, enhance the patients' medical history, and are an excellent resource to drive diagnostic and prognostic medical image analysis. However, these multi-gigabyte morphological representations, that assemble typically small datasets, cannot provide instant visual understanding and impede the direct application of conventional deep learning techniques. The current state of the hardware is unable to facilitate learning from images with such high resolution, and thereby some form of dimensionality reduction is required. Most successful approaches to training deep learning models extract and use patches out of the WSIs applying some sampling scheme. Image patches are usually square regions with the majority of the approaches to use dimensions of 256×256 pixels [42, 57, 62, 113]. Figure 2.1 shows an example of a histopathology WSI of skin data [149] along with close-ups of extracted patch samples.

Figure 2.1: A whole slide image (WSI) of skin tissue (**top**) along with a close-up for better visual inspection and extracted patch image samples (**bottom**). The patches are derived by applying a sampling scheme over the WSI, like a random uniform grid without overlapping.

Computer 3graphics

The term computer graphics describes the use of computers to generate images. Within this discipline, numerous algorithms and mathematical tools are available to synthesize different types of imagery, from realistic to non-realistic representations, visual effects, technical illustrations, fictional computer animations, and more.

The image synthesis pipeline can be divided into two main steps: *modeling* followed by *rendering*. Each step can be conducted in different ways relying on a larger or smaller extent to physically-based definitions or approximations of reality. Synthetic data generation requires to define first a detailed 3D scene containing objects, light sources, materials, camera characteristics, and other possible participating media. The involved geometry consists of surfaces often described by simpler geometric primitives, such as triangles. This step is generally referred to as modeling. It deals with the mathematical specification of shape and appearance properties of a virtual environment in a way that can be stored in a computer. For example, a chair might be described as a set of ordered 3D points along with some interpolation rules to connect them and a reflection model describing how light interacts with the chair.

Light is responsible for anything being visible, from the HVS to cameras, and digital image synthesis. Light transport theory describes the propagation of light emitted from light sources and its interaction with materials on surfaces in a scene. A computer-generated image is then computed by simulating how light propagates in the scene and reaches the virtual viewpoint. This is the rendering step, a term inherited from art, which creates shaded images from the 3D models. The whole process with its consecutive steps is illustrated in Figure 3.1.

This chapter focuses on 3D graphics and the creation of completely synthetic images, from content creation and modeling to the final rendered result. It outlines basic concepts and common practices, provides a better understanding of the physics of light to set the base of light transport simulation and image synthesis, and gives more details for the methods employed by this thesis. For further information and implementation approaches, the reader is referred to some of the rich literature of the field [5, 69, 160, 185].

Figure 3.1: Image generation through a computer graphics pipeline. First, a mathematical model of a scene needs to be defined (modeling step). This includes 3D objects geometry, light sources, material properties, and any atmospheric participating media. Then, the final image is realized by the simulation of the light energy navigating through the scene and reaching the virtual viewpoint/sensor (rendering step). When the light travels from its sources through the 3D scene, it may be reflected several times before reaching the virtual sensor.

3.1 Modeling

Modeling is generally the process of generating the virtual scene, the objects, and the environments in which the sensors are simulated. This step is about developing a representation of all aspects related to the scene content, ranging from the configuration of 3D object models to surface textures and light sources.

Scene modeling has evolved over the years from vector graphics and simple Computer-Aided Design (CAD) systems to open-source or commercial 3D modeling software packages with build-in transformation functionalities, fractal geometry and procedural generation, fluid and other participating media simulations, and many more. Similarly, texture modeling has moved from one-color mode and manual 3D "curving" to create depth, to texture mapping where the textures are created independently of the 3D objects and are then applied and wrapped around the surfaces. Further advances enabled bump (and wrinkles) mapping, as well as structured, algorithmic, and automatic texture generation ranging from noise functions to procedural modeling of various forms. Apart from textures, material representation has progressed through the development and use of numerous illumination models. These can be broadly divided into parametric and data-driven models of the bidirectional reflectance distribution function (BRDF). Extended research on analytical formulations based on physical properties, like roughness or shininess, and physics theory, as well as empirical models fitted to measured data, have enabled material modeling and their photo-realistic representation.

3.1.1 Basics

The geometric models are most often described as groups of triangles with shared vertices. They are known as triangular or simply triangle meshes or triangular irregular networks. The triangles interconnect via shared vertices and edges to represent a distinct continuous surface. This way, a mesh can be collectively and efficiently handled as a network of related triangles. A mesh is practically defined by a list triangles and their respective vertex coordinates, defined at a coordinate system local to the geometry. This is commonly referred to as *modelling* or *object space*. Complex scenes and environments can be created by combining several geometric objects and light sources. The scene is a unique space with its coordinate system, the *world space*, where all models exist and get determined by transformation matrices applied in the vertices and the normals of each model in the respective object space. Thus, each mesh is associated with a model transform that positions and orients it to the world space. Several transforms can be assigned to the same model enabling more than one instance of the same mesh in different locations, orientations, and sizes without needing to repeat the basic geometry. The models visible to the simulated sensor, e.g., camera, are finally rendered. The virtual camera's and lights' location, direction, and characteristics are given in the world space. However, the virtual viewpoint defines its coordinate system too — the *view* or *camera space*. The camera and models are expressed in the view space through view transformations to ease projection and clipping.

Efficient 3D mesh manipulation is critical for the performance of many graphics systems. In this scope, each application defines mesh handling efficiency differently. For example, in computer video games, typical aspects include storage, bandwidth, and operations that modify the triangle mesh structure, like editing, subdivision, or compression, and need fast access to vicinity vertex information. Most similar programs need to also store data at vertices, edges, or faces to support texture mapping, shading, animation, and other operations. Vertex data are the most common. Each vertex can carry parameters that vary across the surface, like material properties, texture coordinates, and irradiances. These values can then be linearly interpolated across each triangle to define a continuous function over the whole surface of the mesh.

The objects in a scene get a more naturalistic appearance through shading, materials representation, and texture mapping. Gouraud and Phong, and the latter's updated version Blinn-Phong, shading are two basic shading methods available in most graphics libraries that provide a matte and more glossy look, respectively [34, 82, 186]. In reality, objects are composed of materials that reflect or absorb light or do both in varying amounts on their surfaces. Considering only BRDF modeling for simplicity, there is a large body of work on parametric and data-driven models that measure how a surface reflects light [17, 50, 156, 242, 269]. Depending on the capabilities of the image synthesis system, accurate appearance modeling is a key feature of realistic representations. Moreover, in practice, hardly any surface is featureless. Materials that may initially have, or are considered having, a clean look can develop at some point marks, dents, stains, scratches, structural details, and other superficial artifacts. These phenomena are not changing the surface fundamentally but rather add some

Figure 3.2: The model of an elephant in all three categories of scene modeling. A physically-based representation next to an also physically-based model of a cat, following the rules of proportion (left). A data-driven approach to learning a statistical model for the shape of an elephant using 3D laser scans (middle). A non-physically-based scene setup involving fictional portrayals and violation of laws of the physical world as we know it (right).

spatially varying effect. Texture mapping enables these features replication in computer graphics. A texture is essentially an image that contains spatial variation and is mapped mathematically to the surface. In image generation pipelines, apart from the main purpose of introducing surface detail, texture maps are used to create shape, shadows, and reflections, illuminate, or even store information.

These basic modeling rules govern any image content creation framework. The following sections describe further common practices and provide more details on the methods used in this thesis.

3.1.2 Common practices

Modeling can be categorized according to the method the attributes of the scene are created. Paper A, which is a comprehensive description of image synthesis techniques and synthetic training data for deep learning with a focus on their computer graphics aspects, provides a general categorization based on three rules. The first one is *physically-based modeling* where the geometry is a result of well-established processes. These are usually expressed mathematically following the laws of physics. The concept is expanded to hand-modeled scenes constructed out of perceptual similarity without using necessarily scientific formulations.

Data-driven modeling involves high-quality sensor data and measurements, like motion capture and laser scanning, to build statistical models for shape formation. This happens by fitting the captured data to parametric models or feeding them to learning-based non-parametric methods. Thus, reliable models of real-world geometries can be shaped, even when there are no available explicit mathematical representations or physically-based equivalents are not possible to be applied. Data-driven approaches have been used widely in the face, body, and material modeling.

Figure 3.3: Automatic urban scene content creation using procedural modeling and shape grammars. Using schemes like this it is possible to generate city street maps, buildings, facades, vegetation and more. The images are retrieved by [174].

Finally, *non-physically-based modeling* includes everything that is not physically-based, meaning any form of scene content that does not follow any physical rule accurately or approximately. Non-physical models are composed on a random, abstract, or fictional basis. Figure 3.2 shows examples of the same concept for all three modeling types to allow for better understanding and comparisons.

3.1.3 Procedural modeling

The constituting elements of a scene can be generated in a non-procedural or procedural manner. Non-procedural modeling relies predominantly on manual efforts and artistic CAD work, and in general does not involve automatic generation procedures. On the contrary, procedural modeling is the use of algorithms and mathematical functions to determine the layout of a scene, the shape of an object, or the color and pattern of a texture. In reality, procedural modeling is about parameterizing the generation process and can be applied to any factor of scene content creation, providing a high level of control, flexibility, and variability.

Figure 3.3 shows examples of procedural modeling of buildings and vegetation for general urban environment content creation [174]. Delving deeper into this concept example, the shaping of the virtual urban world can be defined by a collection of parameters and a rule set that, in turn, maps the parameters into a coherent scene definition. Therefore, each scene arrangement — in essence, each generated image — is a sample of the parameter space. Associating a distribution to each parameter and sampling from them leads to individual unique results. In this scheme, scene content can be defined by fully procedural objects or a mix of them and objects from model libraries. For example, the synthetic data set for street scene parsing, developed in paper **B** and later improved and made publicly available [274], employs procedural modeling for some classes, like buildings and road textures, and pre-made objects drawn from libraries for others, like pedestrians and cars. In this case, increased levels of variability in the shared geometry are possible by diversifying the objects' placement, orientation, textures, and simulated material.

3.2 Rendering

Rendering follows modeling and generates an image out of the 3D scene configuration. It is a computational technique simulating the light transport between light and matter from a defined viewpoint, in various levels of accuracy. It is a fundamental part of computer graphics and a well-studied field, with continuous advances for more than 50 years. Simply, it converts a 3D scene description into a 2D image. It is an interdisciplinary area where several computer graphics processes, like geometric and texture modeling, optics simulation, etc., need to go through some procedure to be visible in the resulting image. The involved algorithms have evolved tremendously over the years for various tasks, from viewpoint visibility to physics simulation.

Light is the essential element for any visual system to create images. The same principle holds for digital image synthesis, where different approximations describe the rules under the light travels from the light sources to the viewpoint and interacts with the participating materials and media in the scene. Each approximation provides a different trade-off between computational complexity, accuracy, and resulting quality. It is, therefore, considered beneficial to briefly explain the physics of light to support a better understanding of how 3D models are rendered to images. The following sections will describe the fundamental concepts of light transport theory and the various approximations used in computer graphics rendering.

3.2.1 Light transport theory

There are four models of light describing its behavior and interaction with surfaces: ray optics, wave optics, electromagnetic optics, and quantum optics [210]. The simplest of these is ray optics, also called geometric optics, which computer graphics typically rely on for *physically-based rendering*. In ray optics, light can only be emitted, reflected, and transmitted while other effects are ignored. Moreover, it is considered to travel in straight lines at an infinite speed.

Radiometry is the relevant scientific framework to quantify light transport providing several measures for electromagnetic radiation. Radiance, noted as $L(x, \omega)$ and measured in watt per steradian per square metre ($W \cdot sr^{-1} \cdot m^{-2}$), is the primary radiometric measure for physically-based rendering, and describes how much light (energy) flows through point x in direction ω . Radiance is most directly related to the perceived brightness of colors to the human eye and is the quantity that needs to be computed for each pixel in a rendered image.

The rendering equation

The *rendering equation* is a mathematical model that describes the interaction of the surfaces in the 3D scene with light [120]. According to this the outgoing radiance $L(x \rightarrow \omega)$ from a point x towards direction ω is equal to the sum of the emitted radiance $L_e(x \rightarrow \omega)$ plus the reflected radiance $L_r(x \rightarrow \omega)$ at the same point and direction. Assuming

that there are no participating media in the scene, such as water, fog or smoke, the reflected radiance can be calculated by integrating the incident radiance $L(x \leftarrow \omega_i)$ at the point x inwards from direction ω_i over the visible hemisphere Ω at x . The rendering equation is then formulated as:

$$L(x \rightarrow \omega) = L_e(x \rightarrow \omega) + \underbrace{\int_{\Omega} L(x \leftarrow \omega_i) \rho(x, \omega, \omega_i) (n_x \cdot \omega_i) d\omega_i}_{L_r(x \rightarrow \omega)} \quad (3.1)$$

where $\rho(x, \omega, \omega_i)$ is the BRDF, and n_x is the surface normal at point x . The BRDF describes the surface reflectance between the incident and outgoing directions, i.e., the amount of incident light coming from direction ω_i that is scattered back towards direction ω [176]. Note that surfaces that are not light sources do not emit radiance, and therefore $L_e(x \rightarrow \omega) = 0$.

The rendering equation models the interaction of light with a scene. In reality, the incident radiance at a surface point is determined by the geometries and scattering properties of all objects in the scene. Image synthesis that accounts for single interaction of light with the scene surfaces is called *direct illumination* and is achieved by applying the rendering equation once. On the other hand, rendering algorithms that attempt to imitate reality's complexity are referred as *global illumination* and take into account not only direct illumination, but also subsequent incidents where the light interacts with the surfaces in the scene multiple times. For this the rendering equation needs to be evaluated recursively.

Modeling participating media

The rendering equation is sufficiently accurate to model optically thin media, like clean air, whereas it is inadequate for cases where the light interacts with denser, participating media in the scene. For these cases, computer graphics, like other scientific fields, consider the medium to consist of countless tiny particles that interact with light, and utilize the *linear transport theory*. The latter is the mathematical framework that describes the interaction of light through any random medium and accounts for a statistical interpretation of the aggregated behavior of the scattering particles that fill a region of space. Similar to other statistical models, one of the main benefits is that there is no need to represent individual particles' position since their average effect is considered and that individual interactions with light can be assumed independent from following events.

The interaction of light with the medium particles is governed by two types of events: *absorption*, which means that the light converts to another kind of energy, and *scattering* to a different direction. The probability of an interaction per unit length is modeled by the extinction coefficient $\sigma_t = \sigma_a + \sigma_s$, where σ_a and σ_s represent the relative probability of a particle being absorbed or scattered per unit length respectively. These probabilities depend on the medium particle density and can vary spatially. Therefore, they define a medium as homogeneous if they remain constant throughout the medium or

heterogeneous if they vary spatially. Both of the particle-light interaction events reduce the radiance along a ray, as the energy can be transformed to other types or scattered into different directions. The scattering to different directions event is also referred to as *out-scattering*, to differentiate it from possible *in-scattering* events where the radiance in the ray is increased as energy coming from other directions can be added along the way. Another source of energy increase can be the *emission* of light in the medium. All possible energy alteration events at a point x towards direction ω are modeled by the integro-differential form of radiative transfer, known as the *radiative transfer equation* (RTE), that states that the change in radiance is the sum of three terms: a decrease in radiance dictated by the extinction coefficient, and an increase due to emission and the in-scattering integral.

$$(\omega \cdot \nabla)L(x \rightarrow \omega) = \underbrace{L_e(x \rightarrow \omega) + L_i(x \rightarrow \omega)}_{\text{energy increase}} - \underbrace{\sigma_a(x)L(x \rightarrow \omega) + \sigma_s(x)L(x \rightarrow \omega)}_{\text{energy extinction}} \quad (3.2)$$

emission in-scattering
absorption out-scattering

where the in-scattering energy increase is given by an integral over the unit sphere S^2 around the position x by:

$$L_i(x \rightarrow \omega) = \sigma_s(x) \int_{S^2} \rho_p(x, \omega, \omega_i) L(x \leftarrow \omega_i) d\omega_i \quad (3.3)$$

where $\rho_p(x, \omega, \omega_i)$ is the phase function that expresses the probability density of light that arrives at x from direction ω_i to be scattered into direction ω , and $L(x \leftarrow \omega_i)$ is the radiance reaching x from direction ω_i .

The RTE can be formulated in an integral form to express the radiance at x coming from direction ω , $L(x \leftarrow \omega)$, using as boundary condition the rendering equation 3.1. It comes more natural to describe where the radiance is coming from with a backward-transport formulation that accounts for all interaction events along the negative part of the radiance beam, instead of the direct forward-transport integration of the RTE. This formulation is generally known as the *volume rendering equation* (VRE) and can be derived by integrating both sides of the RTE 3.2 along the negative direction $-\omega$:

$$L(x \leftarrow \omega) = \int_{t=0}^d \exp\left(-\int_{s=0}^t \sigma_t(x_s) ds\right) \left[\sigma_a(x_t) L_e(x_t \rightarrow -\omega) + \sigma_s(x_t) L_i(x_t \rightarrow -\omega) + L_d(x_d \rightarrow -\omega) \right] dt \quad (3.4)$$

where t and s are the running ray parameters along the negative part of the radiance flow, d the intersection point of the ray with a surface or another medium, $x_t = x - t\omega$, $t \in (0, d)$ and $x_s = x - s\omega$, $s \in (0, t)$ are positions along the ray for the integrals over dt and ds ,

and L_d the incident radiance at the end of the ray. The transmittance from the point of interest x to a point x_t along the ray is modeled by:

$$T(t) = \exp\left(-\int_{s=0}^t \sigma_t(x_s) ds\right) \quad (3.5)$$

that describes the net reduction of radiance due to absorption and out-scattering events between x and x_t per unit distance. The VRE can be then simplified to the following formulation that reveals a reasonably straightforward concept: the radiance arriving at a point from a known direction is equal to the added radiance of all points along the beam to the first surface or another medium, plus the outgoing radiance of the boundary times the ray transmittance.

$$L(x \leftarrow \omega) = \underbrace{\int_{t=0}^d T(t) \left[\sigma_a(x_t) L_e(x_t \rightarrow -\omega) + \sigma_s(x_t) L_i(x_t \rightarrow -\omega) \right] dt}_{\text{radiance from accumulated emission and in-scattering events}} + \underbrace{T(d) L_d(x_d \rightarrow -\omega)}_{\text{radiance from the closest boundary}} \quad (3.6)$$

In computer graphics, the VRE is the equation that must be solved to render images with participating media. It can be perceived as a generalization of the rendering equation that enables volume phenomena representation. In this context, surfaces can be perceived as very dense volumes with the appropriate phase functions. It is, however, a costly approximation solution system, as the increased dimensionality leads to increased complexity.

3.2.2 Light transport simulation

Simulating light transport has been a central topic in computer graphics as it is the primary element for photo-realistic results. The RTE (3.2) is a Fredholm integral equation of the second kind, with no analytical solution, apart from the trivial cases. Therefore, different solutions have been developed to approximate it, with varying levels of complexity. One example of the most accurate approximations is the family of stochastic Monte Carlo methods. These are numerical methods based on averaging many random samples to calculate estimates of the true solution [136, 185]. Another example of approximating the solution for light transport in a scene is the class of deterministic methods. Radiosity is a finite-element approach that computes the illumination distribution by simulating the energy exchange between discrete small surface patches, or participating media voxels, in the scene [81].

This thesis focuses on surface light transport and utilizes the rendering equation and its approximation solutions. The rendering equation expresses the amount of radiance that originates from the light sources, scatters at surfaces in a scene, and finally reaches the camera. Image synthesis is performed by estimating the steady-state equilibrium of

the rendering equation (3.1). This requires solving it for a large number of sample points in the image plane, potentially millions, ending up with equally many inter-dependent, high-dimensional analytically intractable integral equations. This is particularly challenging since the outgoing radiance $L(x \rightarrow \omega)$, that the equation is solving for, also appears inside the integral formulation (3.1). The reason is that the outgoing radiance from any point impacts the incident radiance at every other point in the scene. As a result, the rendering equation becomes a massive system of nested integrals. In practice, the rendering problem can be solved using numerical integration, which can be conducted with different approximations. The rendering computations are bound to the available, at the time, hardware capacity and the critical for the end visual application usage and resources. There is a wide range of application types, each with its own set of requirements. According to them, rendering can be broadly categorized into two main groups: *real-time* or *online*, that is generally performed on the graphics processing unit (GPU), and *offline* that performs computations on the central processing unit (CPU) too.

Each category holds its pros and cons, fulfilling different kinds of objectives in a data generation pipeline. For example, real-time rendering has the benefit of fast image generation but is restricted to light pre-computations to achieve realistic results that do not scale well in large-scale, content-diverse image production. On the other hand, offline rendering can simulate any kind of light transport and scattering effect, with a high degree of flexibility and scalability but at the cost of higher rendering times per frame. Figure 3.4 shows a comparison between real-time and offline rendering examples.

Real-time rendering

Real-time rendering produces images with high frame rate for immediate user interaction, such as in computer video games. Therefore, the critical resources include the processor time, memory, and bandwidth associated with it. The GPU advances over the past years marked the transition from the fixed-function pipeline to programmable graphics and rasterization rendering. In this method, the 3D objects — represented by meshes of polygons, e.g. triangles — are traversed on the screen and converted into color pixels. Realistic results that include complex materials and lighting effects can be achieved by associating normals, textures, and colors to the polygons and using pre-computation for light simulations. Most computer game engines use this technique. It is generally very fast, allowed by GPU computations and quick generation of continuous-time instances in the same environment. Pre-computations enable complex scattering effects or global illumination (light rays have bounced two or more times at surfaces in the scene), delivering more accurate sensor simulation and approximations of the rendering equation. However, they add up to the computational complexity and may cause scalability problems. Today, modern GPU vendors provide powerful hardware that can simulate higher-quality approximations of the physics of light-object interaction in real-time.

Moreover, ray-casting is a visibility algorithm that can be used in real-time. The idea

behind it is to shoot rays from the viewpoint, one per pixel, and find the closest object blocking the path of that ray [14]. Then, the image pixel color is calculated for the hit point according to the radiance emanating from the visible surfaces.

Offline rendering

Offline rendering is used when the image generation is not tied to specific time frames. Typical examples include visual effects or fully computer-generated movies. This family of methods can accurately simulate the complex physics-based algorithms that directly model the intrinsically recursive nature of inter-object reflections. Ray-tracing, the successor of ray-casting, shoots rays into the scene and resolves for reflections and refractions to get the pixel colors [271]. Ray-tracing is used by rendering techniques that account for global illumination, apart from direct lighting, like path-tracing [120], which involves Monte Carlo integration to simulate complex lighting effects, and photon mapping [115, 116]. Path-tracing techniques, used in paper **B**, can accurately simulate the light distribution in a 3D scene by calculating the energy that travels in the rays and paths connecting the viewpoint, light sources, participating media, and objects. These methods can simulate any form of light transport and a broad range of sensors without any pre-computation needed. As a result, they can render any lighting effect, including multiple bounces and combinations between complex geometries and material scattering responses. It is also possible to sample the scene over the spatial and temporal domain. Several paths per pixel sample the rendered area over the extent of one pixel, which is typically the case on a real camera sensor. This procedure acts as an anti-aliasing algorithm and permits the simulation of optical systems' point spread functions (PSF). Additionally, the path samples can be distributed over time, making it feasible to simulate motion blur. This generality is many times desirable by visual data synthesis for deep learning. It is often essential to generate large numbers of diverse image samples. This need for diversity, i.e., content variation, makes the use of light pre-computations impossible due to their lack of scalability capacity. Path-tracing is a standard tool in film production, implemented in many rendering engines. The impediment is the high computational complexity inherent to solving the light transport. However, these techniques parallelize well, and it is common to employ high-performance computing data centers and cloud services for large-scale image production.

Along with radiosity, these methods underlie what is commonly known as *physically-based rendering*. To date, offline physically-based rendering is most often the only way to achieve photo-realism.

(a) Real-time rendering example retrieved by SYNTHIA [203]. The dataset is generated in a 3D development platform employing non-procedural physically-based modeling.

(b) Offline rendering example retrieved by Synscapes [274] (the successor of paper B). The dataset employs a physically-based procedural modeling engine coupled with path tracing and Monte Carlo integration.

Figure 3.4: Real-time (a) and offline (b) rendering examples, retrieved by two publicly available datasets for street scene parsing. Note the differences both in scene geometry accuracy and the achieved degree of realism (shading, reflections, highlights, shadows, and textures fidelity).

Generative modeling 4

AI strives to make machines/computers analyze and “*understand*” common and complex notions of the world, like visual representations. As part of this, machine intelligence research has developed generative modeling which aims to describe several concepts, e.g., image synthesis for computer vision or sound synthesis for natural language processing, in terms of probabilities and statistics. Generative image modeling refers to the collection of the statistical and probabilistic methods that learn visual representations out of given data distributions and can synthesize new samples out of them, without any prior structural knowledge, as computer graphics require. Due to their capabilities, a wide range of models exists, mainly classified in autoregressive networks, variational auto-encoders (VAEs) and generative adversarial networks (GANs).

Deep generative image models are generative models that are parameterized with deep neural networks, providing in this way higher levels of scalability and high-dimensional image data generation capabilities. Although relatively recent, the most representative and widely used example is GANs, which are inspired by game theory and based on a minimax game between two adversary neural networks. Today, GANs are considered the state-of-the-art generative models for image synthesis producing highly realistic, convincing synthetic images.

This chapter focuses on generative image modeling, and GANs in particular. To build the appropriate theoretical foundation, it first outlines fundamental concepts related to the possible representations of the probability density function. Then, it describes in detail the basic GAN framework, providing its strengths and challenges, and finally gives enriched information for the GAN variants used in this thesis to generate synthetic image data for deep learning. For supplementary descriptions and analyses, the reader is directed to additional literature of the plethora available [103, 180, 215, 268].

4.1 Fundamentals

It is established that generative modeling constitutes a powerful image synthesis method that reduces the dependency on prior knowledge, compared to computer graphics, as it learns and generates new imagery from observed data. In this family of methods, density estimation is a central factor. Essentially, in density estimation, the generative model learns to estimate the unseen probability density function of the original data. On that basis, generative models can be grouped into two general categories according to how they represent the probability distribution, either explicit or implicit.

The first one is **explicit** or **maximum likelihood-based generative models**. The basic idea of maximum likelihood is to define a model that provides an estimate of a probability distribution, parameterized by a set of parameters, e.g., θ . Then the likelihood is the probability that the model assigns to the training data, i.e., $\prod_{i=1}^n p_{model}(\mathbf{x}^{(i)}; \theta)$, for n training set observations. The principle of maximum likelihood optimizes for the parameters θ that maximize the likelihood of the training data. Intuitively this means that in a hypothetically precise optimization process the true density distribution p_{data} would lie within the family of distributions $p_{model}(\mathbf{x}; \theta)$. In practice, this is rarely the case. Moreover, as in the log space a product converts to a sum, which in turn can further lead to simplifications of mathematical expressions and digital computer implementations, it is common to maximize the log-likelihood of the training set [31, 79].

Likelihood-based generative image modeling is grouped into models that guarantee computationally tractable density probability and models that admit tractable approximations to the likelihood and its gradients. Typical tractable explicit models are the family of autoregressive models, such as PixelCNN [179, 212], PixelRNN [258], the masked auto encoder for distribution estimation (MADE) [76], and the neural autoregressive distribution estimator (NADE) [138, 257]. In the same category also fall normalizing flow models that define continuous, nonlinear transformations between two different spaces. In this type, complex distributions are defined by transforming a probability density through a series of invertible mappings [198]. Typical representative models used for image synthesis include non-linear independent components estimation (NICE) [58], real-valued non-volume preserving transformations (real-NVP) [59], variational inference with inverse autoregressive flow (IAF) [132] and masked autoregressive flow (MAF) [181]. Models that define explicit tractable probability densities are highly effective as they allow for optimization directly on the log-likelihood of the training data. They are, however, limited and slow for high resolution and quality image generation. This brings us to the models with explicit intractable density functions that require approximations to maximize the likelihood. These are usually deterministic approximations using variational methods or stochastic approximations using Markov chain Monte Carlo methods (or a combination of both [209]). Variational auto-encoders (VAEs) are the most commonly used network in variational learning in generative image modeling [131, 199], while Boltzman machines are models that use Markov chains to train or to generate samples from the model [4, 66]. Their inflexibility to scale to high dimensional spaces and their computational cost make them less popular than other

methods.

The second general category of generative models refers to the cases trained without defining a probability density function, known as **implicit generative models**. Here, the probability distribution is implicitly represented by a model of its sampling process, i.e., the model is trained while interacting indirectly with p_{model} by sampling from it. Some generative models of this category, called generative stochastic networks [25], use the Markov chain as a stochastic process to transform an existing sample to obtain another one from the same distribution. Other models can generate new data in a single step without any starting point, like moment matching approaches [147] and GANs [78] where new samples from the data distribution are synthesized by transforming, e.g., a random Gaussian vector with a neural network. The basic idea behind GANs is to put two adversary differential functions through a minimax game.

Today, GANs, along with variational autoencoders and autoregressive models, are one of the most popular approaches to deep generative image modeling. In the past few years, their advances are considered the state-of-the-art algorithm for image generation in this family of techniques. It is one of the most active research fields where the developed models usually combine different concepts to achieve stable training on high-resolution data and can produce very convincing synthetic images [39, 124, 126]. This thesis explores image synthesis for deep learning applications using mainly a GAN variant that can translate/transform images from one domain to another.

4.2 Deep generative models

Generative models — both of the explicit and implicit families — have been parameterized using deep neural networks, that combined with the advances in stochastic optimization methods, formulate the category of *deep generative models*. These are scalable methods able to model high-dimensional image data and have been demonstrated on belief networks, Boltzmann machines, VAEs, GANs, autoregressive models, normalizing flows, energy-based models, and score-based models. The application areas that benefit from deep generative models include image synthesis and computer vision, speech and natural language processing, graph mining, reinforcement learning, and more.

In visual data generation, the various kinds of deep generative models provide different levels of efficiency and efficacy, with GANs being the most widespread solution. Markov chain Monte Carlo-based explicit approaches calculate the log-likelihood gradient where gradients vanish during training. As a result, image generation becomes a slow procedure, as it is not possible to mix between modes fast enough. VAEs use deep learning with statistical inference to represent a data point in a latent space and, as an explicit but intractable density probability approach, undergo the complexity of approximations. They are trained to maximize the training data likelihood and, therefore, they can experience the curse of dimensionality in many image sets. Moreover, sampling from the Markov chain in high-dimensional spaces can be computationally slow,

as mentioned, blurry, and inaccurate. PixelCNN and PixelRNN deep autoregressive models translate the problem of image generation into pixel prediction and generation. Thus, they rely on a per-pixel procedure, when GANs can generate an image sample in one shot and faster times. Moreover, by default, the implicit nature of GANs, which do not require a density probability and come with an internal adversarial training mechanism, provides more flexibility than implicit model alternatives.

4.3 Generative adversarial networks

GANs can handle most of the restrictions of the rest of the standard deep generative models that were mentioned above. Although they are a relatively new approach, they represent a widespread way of generating new image samples with impressive results. Inspired by game theory, they constitute a novel class of deep generative models in which back-propagation is used for training to evade the issues associated with Markov chain Monte Carlo training and the difficulties related to maximum likelihood learning.

In contrast to computer graphics methods, GANs operate directly in pixel space and produce an image out of a network. They are ordered probabilistic models that incorporate two adversarial deep neural networks. The generator (G) is trained to capture the true data distribution, while the discriminator (D) estimates the probability of whether an image comes from the real or generator's data distribution.

The generator input is a random noise vector z (usually from a uniform or normal distribution $p_z(z)$). This is mapped to a new data space creating a fake sample $G(z)$. The discriminator is a binary classifier that takes both a real sample x from the true distribution p_{data} and the fake image generated by the generator and determines if it is real or fake. The discriminator and the generator have opposite objectives, i.e., the discriminator strives to minimize $D(G(z))$ and prove the generated sample as fake, while the generator aims to maximize the same quantity. The loss function of the discriminator can be expressed as

$$\mathcal{L}_D = -\log(D(x)) - \log(1 - D(G(z))), \quad (4.1)$$

aiming to determine the source of the data correctly. $D(x)$ and $D(G(z))$ describe the probability that x is discriminated as real data, or as a sample generated by the generator (fake), respectively. The discriminator is then optimizing for $D(G(z))$ close 0, while the generator wants to bring it closer to 1. The generator's loss function is

$$\mathcal{L}_G = -\log(D(G(z))). \quad (4.2)$$

During the optimization process, the two loss functions are iterated, and the problem is transformed into the adversarial minimax game as shown below,

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))], \quad (4.3)$$

with value function $V(D, G)$. Generator and discriminator compete in a minimax zero-sum game until Nash equilibrium is reached using a gradient-based optimization technique (simultaneous gradient descent). The generator and discriminator parameters are

Figure 4.1: The basic architecture of generative adversarial networks (GANs). Two deep neural networks, the generator (G) and discriminator (D), are synchronously trained during the learning stage. The generator's input is a random noise vector z (usually from a uniform or normal distribution) that is mapped to a new domain creating a fake sample $G(z)$. The discriminator is a binary classifier that takes both a real sample x from the true distribution and the fake image generated by the generator and determines if it is real or fake. It is optimized to distinguish between real and generated images while the generator is trained to fool the discriminator from discerning between real and fake samples.

updated iteratively. Gradient signals are received from the loss induced by calculating divergences between two distributions by the discriminator. The architecture of GANs is illustrated in Figure 4.1.

When the model is trained, the generator creates data that looks real enough that the best the discriminator can do is guess randomly. This means that during training, the generator learns to improve the synthesized images by adding more real-looking features, but at the same time, the discriminator also learns to be more efficient and recognize more fake features. This way, the training loop forces the generated data distribution $p_{G(z)}$ to converge to the true data distribution p_{data} .

4.3.1 Challenges

As in every deep neural network, hyperparameters selection, such as the learning rate, batch size, momentum, and weight decay is of utmost significance for GANs training convergence [191].

The minimax game in the vanilla formulation described above, is sensitive and prone to fail. The generator and discriminator may oscillate during training instead of converging towards a fixed-point. If one model gets more powerful and takes over, the gradients for optimization can vanish, leaving the stochastic gradient descent ineffective and the learning process unstable. For example, when the generator produces poor samples, the discriminator learns to differentiate them easily from real images. Then the probability of the generated images to be classified as real is close to 0, i.e., $D(G(z)) \approx 0$ and, therefore, $\log(1 - D(G(z))) \approx 0$, too, and the discriminator fails to provide gradients. In turn, this means that the generator stops updating, and the entire system gets unstable.

Another problem is *mode collapse* and *drop* [278], where the generator focuses on one or a few modes respectively and cannot represent the diversity contained in p_{data} . In these, the generator fails to produce more than a single type of training data, or produces just a few of them, even though it might be a rich image set. These modes indicate a faulty training procedure caused, for example, when the generator finds a data type that can easily fool the discriminator and, therefore, keeps generating it. Oscillations like these impede reaching Nash equilibrium and lead to unstable GAN training.

Even though GANs are widely applied to various tasks, often with visually pleasing results, it is still not universally defined how to quantitatively evaluate them. Models that generate “good” samples can have poor maximum likelihood, i.e., a low probability that the model is similar to p_{data} , and models with better likelihood can generate “bad” samples, while they may also differ in model formulation and training and evaluation practices. Their assessment is still, to a large extent, qualitative — through human visual inspection — although some approaches and measures have been introduced to quantify performance. Visual examination is time-consuming, subjective, and cannot capture distributional characteristics. Similarly to the cruciality of the appropriate model selection to achieve the best possible performance for a task, the suitable evaluation metric selection is necessary to draw the right conclusions. With the emergence of new models, some metrics have been introduced to fulfill the analysis needs of different tasks. These metrics aim to measure the quality of the GAN-generated images and include the *inception score* (IS) [211], *mode score* (MS) [177], *fréchet inception distance* (FID) [96], *kernel inception distance* (KID) [32], and multi-scale structural similarity for image quality (MS-SSIM) [178, 267]. However, not all of them are suitable in all cases and true indicators of potential problems. For example, the IS, defined to scale image quality and diversity, fails to highlight mode collapse incidents, an issue that is addressed by the MS metric. Moreover, both the IS and FID have problems in capturing the occurrence of overfitting, which KID tries to tackle. Apart from the specifically defined metrics, they have also been used binary classifiers to determine whether different samples come from the same distribution [144, 155], and loss functions to measure the distance between the true and generated imagery distributions [16, 83].

A key challenge is that it in many cases is difficult to define appropriate evaluation metrics. Further research and development is needed to establish quantitative metrics and techniques that can accurately distinguish between real and generated samples,

verify mode collapse, mode drop, and detect overfitting. A solid foundation for further GANs advancements mandates to overcome the limitations of the qualitative metrics and establish quantitative ones able to distinguish the generated samples from the real ones, verify mode collapse, mode drop, and detect overfitting. These, along with possible meta-measures, can provide researchers with a complete guide and the right tools for accurate task evaluation [35].

4.3.2 Common variants

Since their debut in 2014, a large body of work has been devoted to improving GANs' training stability and output image quality. Given that the main components in the design and optimization of GANs are, besides the optimization algorithm, the network architecture, and loss function, there are two corresponding variants of GANs. Since GANs' building blocks are neural networks, naturally, specific architectures may be beneficial for specific applications, e.g., CNNs for images and RNNs for time series data. Loss-variants explore different variations of the loss function to enable more stable generator learning.

Architecture-focused GAN variants also include flavors that focus and modify the latent space and versions that perform alterations depending on the target application. Therefore, this category refers to variants that improve the overall network architecture or present modifications based on latent space representations or according to the needs of an application family. On the other hand, loss-variants focus on the various types and per-objective alterations of the loss function and apply/develop different kinds of regularization techniques to further perform penalization using the loss function or other types of network normalization.

Examples of architecture-variants are the deep convolutional GAN (DCGAN) that focuses purely on the network aspect and was the first to provide a stable strategy for image generation with GANs in a single shot process [192], progressive growing GAN (PGGAN) that involves progressive steps toward the expansion of the network [124], the latent space conditional GAN (CGAN) that extends the vanilla formulation by conditioning both the generator and discriminator on extra information [169], and the application-focused couples of GANs [150], cycle-consistent GAN (CycleGAN) [294] and style-GAN [125] that provide image generation alternatives based on image-to-image translation and advanced style-transfer controls. Moreover, widely used loss-variant representatives are the wasserstein GAN that allowed for more robust optimization [16], and f-GAN that provides a unified framework based on f -divergence [177]. For further details on recent advances, a thorough review and analysis are provided by Pan et al. [180] while Wang et al. classify the various variants collectively under computer vision [268].

This thesis utilizes architecture-variants that focus on specific applications or are based on neural networks variations. Image synthesis is approached through unpaired image-to-image translation and noise-to-image generation. For this, the models employed in the attached papers are briefly described.

Figure 4.2: Forward cycle of the CycleGAN variant. In this, one real sample from domain X is mapped/translated to another Y through an encoder-decoder generator. The associated discriminator D_Y determines if the generated sample in domain Y is real or fake. The cycle-consistency loss intends to regularize further the mapping by intuitively trying to bring closer the reverse mapping of the fake sample, i.e., the reconstructed image in domain X, and the original sample in domain X. Backward cycle is similar for domain Y real samples. The depicted image samples are retrieved by [294].

Cycle-consistent GAN — CycleGAN

Image-to-image translation converts the image content from one domain (X) to another (Y). It is described by a family of methods where the objective is to learn the mapping between the input and output images using a training set of aligned image pairs. Even though the proposed methods, which build upon CGANs, have proved to be effective in graphics and vision tasks, they require the training space to be strictly paired (image from X with a corresponding in Y) [111, 263]. In practice, such paired data are hard to find. CycleGANs solve this problem by performing image-to-image translation on unpaired training data.

Thus, CycleGAN is classified as an application-focused variation. To achieve the unpaired image-to-image mapping the generator employs an encoder-decoder framework while, additionally, cyclic consistency is adopted to enforce the translation of unmatched data from domain X to Y. Specifically, the model is based on two mapping functions, from domain X to Y (forward cycle) and vice versa from Y to X (backward cycle). These are paired with two adversarial discriminators, associated with domain Y for the forward (D_Y) and X for the backward cycle (D_X) that encourage the generators to translate one domain to the other producing indistinguishable results. The cyclic consistency loss is introduced in each pass to regularize the mapping. It builds on the assumption

that if one domain is translated to another, then the reverse mapping should return to where it started. Figure 4.2 shows the forward cycle, while a similar analogy applies to the backward one too.

Similar to CycleGANs, more methods that solve for domain adaptation with unpaired training data over two or more domains have been also proposed [47, 129, 282].

Deep Convolutional GAN — DCGAN

DCGAN represents one of the pivotal moments in the history of GAN image synthesis as it was the first one to enable high-resolution image generation. It is a convolution-based variant where the generator utilizes a deconvolutional neural network, inspired by feature visualization for CNNs [287].

Higher image resolution is enabled by leveraging the up-sampling capacity of deconvolution in the generator. Moreover, this variant has three key deviations from the vanilla fully-connected approach that benefit high-resolution modeling and training stability. Specifically, the use of any pooling layer is eliminated, and fractional-strided convolutions and strided convolutions take their place in the generator and discriminator, respectively. Batch normalization is also employed for both of the constructing networks, helping to keep similar statistics for both the generated and real samples. Finally, ReLU and LeakyReLU are used to activate all layers in the generator (except for the output layer) and discriminator, respectively, enabling faster learning and preventing the “dead” neuron effect. The reason is that LeakyReLU allows for small gradients when a unit is not active, and the generator receives gradients from the discriminator.

Progressive growing GAN — PGGAN

Inspired by progressive neural networks [208], PGGAN is also an architecture-variant that involves progressive steps when training the network. The main benefit of this gradual expansion of the model is that it has a form of “memory” that enables the deployment of prior knowledge via lateral connections to previously learned features. This quality has made it quite popular in learning complex tasks resulting in realistically plausible synthetic images [39, 124, 125].

Training is initiated with low-resolution images. Subsequently, more layers are incrementally added to both generator and discriminator. This gradual network growth enables a progressively increasing resolution leading to a higher resolution of the output images, while allowing more stable learning for both generator and discriminator. Intuitively, by increasing the resolution step by step, the networks are called to solve a simpler problem per layer, compared to the end goal of defining a mapping from latent vectors. During the training process, all variables remain trainable. LeakyReLU is used for all layers in both networks (except for the generator’s last layer), and only pixel-wise normalization is deployed in the generator, with no further normalization techniques (batch, layer, weight normalization) in either of the networks.

5 Synthetic data for deep learning

This thesis approaches deep learning-based application research and development from a *data-centric* perspective. In this, the central element is visual data. The model configuration is held fixed while the involved data are systematically improved. Figure 5.1 shows the typical pipeline. The cycle highlights that data is an essential ingredient that enables the models to achieve their expected potential and advocates for their systematic improvement to advance the application's accuracy, robustness, and generalization. For this, it relies on iterative modifications and revisions of the underlying data.

The rise of this direction is relevant but not limited to the increase of training data sizes, which is one of the main reasons for deep learning modern evolution. Advances to the learning algorithms and models have also been crucial. However, they have not been as radical as the progress of training data availability, size, and sophistication, which are all, in turn, a direct consequence of the increasing digitization of society and networking. Apart from visual data collection, other common data-centric practices include data generation, annotation cleaning and consistency, augmentations, documentation, and data debt and governance. Studying and applying these principles to real/captured data reveal vulnerabilities and implementation difficulties. Image synthesis designed for supervised deep learning has the potential to accommodate many of the involved challenges and address the risks, by providing well-defined data with automatic labels.

The presented work contributes to two aspects: data generation for training deep models and synthetic data for augmentation purposes. In this framework, both computer graphics techniques, that generate synthetic training data of natural images for semantic segmentation (paper **B**), and generative modeling, to create anonymized histopathology data for breast cancer classification (paper **F**), were used. For data augmentation, generative adversarial networks (GANs) were employed. The showcased applicability relates to both images from the medical domain (papers **C** and **D**) and natural images for weather classification (paper **E**).

This chapter focuses on these contributions and places them in the context of data-centric AI. It aims to walk the reader through the data-centered practices of a deep learning-based visual application and emphasize the conducted work with specific descriptions, analyses, and discussions.

Figure 5.1: Typical pipeline of a *data-centric* deep learning application, from task definition to deployment. Training images are collected or synthesized in compliance with the defined downstream task. These, with their corresponding ground truth labels, are used to train the specified neural network. The trained model is then integrated into research or commercial production environments and used in practical decision-making processes. The data-centric cycle advocates for systematic improvement of data to advance the application’s performance. It is based on iterative data modifications and revisions to achieve high levels of accuracy, robustness, and generalization. It has to be noted that the evaluation, error analysis, and training data revision process should not be performed based directly on any pre-defined application-specific test set. A separate validation set should be used to avoid overfitting specific images and plasmatic generalization.

5.1 Data-centric AI

Machine learning systems are iterative processes that turn empirical data into usable models using computational algorithms. Deep learning has been, so far, in its vast majority of practices *model-centric*, meaning that the development focuses on building good enough models to deal with the collected observations. In a model-centric view, the system development iterates over network and algorithm tuning to improve performance while the training data are held fixed to the form in which they were collected. On the other hand, *data-centric AI* represents the practices that focus instead on the underlying data, used to train and evaluate the model, to improve performance.

The transition to data-centric approaches is recent. These practices do not aim to downgrade the importance of model development but rather to highlight the neglected potential of improving the performance of AI systems by systematically improving the data

involved. This should not be confused and associated only with the increasing training data sizes. “*Big Data*” has made machine learning more practical and is responsible to a large extent for the rise of the third wave of deep learning. Until then, machine learning systems were called to generalize to a target domain after observing a small amount of data. Big data has lessened this vital obstacle of statistical estimation. Although using larger datasets has been critical to machine and deep learning success, the effort of building appropriately sized image sets is demanding, especially for supervised learning, where careful and thorough collection and annotation are required.

Dataset size is important but not the only element to carefully consider in a visual data-centric deep learning system. A collected/generated training set does not necessarily need to be huge. A high-quality image set needs to incorporate all the features important to the downstream task and can further utilize techniques to generalize to the target domain. Data quality is paramount and multidimensional. It refers to labels consistency and data coverage, accounts for domain shift, and respects individuals’ rights, ethical and societal norms, and legal regulations.

An appropriate image label definition is unambiguous and should remain consistent throughout the samples of the entire set, especially for smaller datasets due to the limited representations. Noise-free labeled data are necessary to make the most out of the observations at hand and can directly affect the model’s performance. Data coverage and the reduction of the related domain shift between the training and testing sets are further central aspects of data quality concerning large and small datasets. The images used to train models must cover the whole spectrum of the expected distribution, including sufficient examples of both common and rare long-tail events. Although long-tail events exhibit a low probability of occurrence, they often capture events with high impact on the application performance, e.g., unusual traffic circumstances in self-driving cars, or occluded objects in detection and classification. Adequate data coverage helps bridge the gap between the training and testing data. However, deficient data coverage is not the only reason for the commonly observed domain shift. The capturing or generation process can introduce discrepancies as well. It is, therefore, necessary to quantify the deviations between the training and testing domains and revise the training data accordingly to ensure reliable deployment. Last but not least, high-quality, “good” data account for fairness, bias, and privacy whenever applicable. Even though this aspect gained more visibility only in the past few years [30, 53, 216], the emerging learning technologies in the broad strata of society worldwide impose the awareness and responsibility by all the involved practitioners.

The core value of data-centric AI, i.e., the systematic improvement of data quality, seeks to enhance the end application performance but also to allow more models to succeed. It is about formulating and providing the tools and techniques to engineer, or *manipulate*, the data to boost the model’s potential. It inverts the traditional perspective by keeping the model fixed and iteratively adapting the training sets, highlighting, the importance, and impact of maximum leverage of data. Practically, the model-centric and data-centric approaches are not antagonistic but synergistic. Deep learning models development

and deployment should include both practices balanced on the downstream task and application needs.

5.1.1 Common practices

Data-centric deep learning strategies tackle practical data problems and often require interdisciplinary methods. Common issues involve data collection/generation, consistent annotation, cleaning, augmentation, documentation, debt, and governance. Some of them are well developed, in contrast to others that are still nascent. While there are no best strategies aggregated in engineering tools, weaving usual practices together into an interlacing whole to provide simple, cheap, and methodic ways of data fabrication, curation, preprocessing, and evaluation will elevate data-centric AI applicability.

This thesis covers aspects of data generation (chapters 3 and 4, and section 5.3), in comparison to data collection (section 5.2), as well as data augmentation (section 5.4). However, for completeness, brief descriptions on the rest of the aforementioned practical problems are given.

Training *data cleaning* identifies the instances whose labels are, or are likely to be, corrupted. Dataset pruning and label noise-cleaning are methods to this end [8]. In dataset pruning, the images with corrupted labels are excluded. While this results in loss of information — and is viable only if the available training set is sufficient in size and diversity — preventing the harmful impact of noise may result in better performance. However, there is the risk of unnecessary data loss by removing too many samples (over-cleaning). An alternative is to remove just the labels of noisy data and reformulate the dataset into a labeled and an unlabeled subset. Then, semi-supervised learning algorithms can be employed. Label noise-cleaning avoids data loss and builds instead upon correcting the suspicious labels to their corresponding true values. Cleaning whole datasets manually, even partially for the noisy samples, is costly and not scalable. Thus, various label cleaning algorithms have been developed. Most rely on feature extraction to map the data to a feature domain that can investigate the noise. This can be done in a data preprocessing fashion, using pre-trained networks, or in a dynamically evolving iterative process based on the model under development: as it gets better, label cleaning gets more accurate, and as label quality improves, the model gets better. Similar to data pruning, these methods can suffer from over-cleaning that is usually associated with misinterpretation of informative hard samples as noisy data or label correction in inappropriate ways.

The goal of *data documentation* is to systematically record information about datasets' purpose, composition, collection or generation process, preprocessing, intended and recommended use, availability, and maintenance [46, 73, 101, 167]. This kind of collection shall also include roles and characteristics of the involved participants, e.g., funding providers, data workers, and data subjects, and the practices and tools that have been used. This has been proposed to strengthen the dataset's credibility, accountability, and transparency to possible fairness, privacy, and consent issues. Moreover,

dataset documentation with formal definitions and instructions for annotations can enhance labels consistency, a central concept to data-centric AI. Most prominently, data documentation aims to set the ground for the dataset creators and dataset consumers' communication [73]. This way, deep learning practitioners in need of training data can make informed decisions and select the appropriate dataset according to their objectives and standards. At the same time, training data collection/generation can be beneficial for noticing about and improving weak spots, e.g., labels, privacy leaks, or strong biases, as the creators are encouraged to record and justify their practices, given that they maintain transparency in all stages. Finally, data documentation, put in the academic context, can contribute to research soundness and correctness, reproducibility, and open science [75].

Data debt is part of the high cost of the so-called technical debt in machine learning systems [217, 218]. Technical debt is a term borrowed from software engineering that accounts for the long-term costs of a deployed model incurred by moving “*fast*” under development. The most dangerous kind of debt is the hidden one, as it builds and scales up silently. Data debt is hidden and mainly refers to data dependencies that have proved to be more expensive than the deep learning code related debts. Data dependencies are a multifactorial issue that is highly related to data selection, annotation, and verification. The model depends on the training images, and these, in turn, need to be stable in terms of quality, quantity, and ownership. Additionally, the training data distribution should care for features that may have a small incremental benefit to the model but make it unnecessarily vulnerable to changes. Regular data versioning and feature evaluations are example tools to control and prevent large data dependency chains. The lack of data documentation and the deprecation of “*old*” datasets is part of data debt. However, the neglected training sets remain in use through, e.g., transfer learning or practitioners that have kept and keep using them. This ephemerality, caused by the constant demand for new and more novel models, does not only increase the technical debt, but can also block scientific reproducibility [216].

Finally, *data governance* in the general context of decision-making AI focuses on the mechanisms of authority and control of not only the used images but also the systems that collect/generate, manage and use them. This is usually expanded to several persons or multi-organizational level, meaning that trusted frameworks are required to ensure reliable data-sharing. Well-defined data governance is necessary to establish accountability and can contribute to transparency and fairness [108]. Part of its mission is that the *right* data are shared securely and reliably to the involved parties while complying with the general data protecting regulation (GDPR) [183] and other relevant laws and regulations. Data governance is one of the success factors for deep learning-based computer vision applications. It increases the value of data and minimizes data-related costs and risks, especially when the data concern individuals, explicitly or implicitly, in the form of natural images [53] or personal health information [92, 273]. Simple as it may sound, and in cases even distant from the academic research framework, data-related mistakes on visual AI applications can lead to systemic bias, violations of individuals' privacy and rights, unlawful decisions, financial and political implications, lives lost,

or any combination thereof. *“In the interconnected world, where data is collected by (and about) governments, businesses, and citizens, and is processed by different entities using various algorithms, dependencies grow, mistakes accumulate, and accountability is gradually lost in the process”* [114].

5.2 Data collection

Capturing visual data, or rather data suitable to solve computer vision tasks, i.e., digital photographs, light detection and ranging (LIDAR) data, scans of whole slide images (WSI) of histopathology data, and more, involve some type of physical sensor and the appropriate hardware and software processes. These are well-established and beyond the scope of the thesis to describe in detail. What it is interesting to mention, though, is the various sources where training data for visual machine learning are nowadays coming from, apart from application-specific targeted capturing and annotation. In the era of increased digitization of society, where many activities take place on computers and can be recorded, while the computers are increasingly networked together, the predominant data collection method is to centralize shared over-the-web visual records and curate them into datasets for deep learning applications.

Thus, formulating image sets of tens of thousands to tens of millions of examples was enabled since the mid-2010s. There are several datasets of importance in the history of machine learning development and specifically computer vision. Among the plethora of the collected visual sets, *CIFAR-10/100* are two of the first large-scale and most widely used datasets consisting of 60K images and covering 10 and 100 classes respectively [134]. They are subsets of the equally widespread *Tiny-Images* set [243] that contain 80M low-resolution images in 76K classes and has recently formally withdrawn due to the existence of categories with derogatory terms and offensive representations. *JFT-300M* [48, 97] is an internal Google dataset of more than 300M images across 18K categories. *Open Images* [137] is its open-source subset consisting of 9M images across 20K categories. Most importantly, *ImageNet*, and its various versions, is the most celebrated and now canonical large-scale image dataset [55, 207]. It includes more than 14M images in total, over 22K different categories, and even though new sets have been introduced since then, it remains the most dominant image database.

The collection process can, for example, rely on some word bank, which is used to automatically download images of the defined terms through some internet search engine. Then, the collected visual set may be manually or semi-manually annotated, i.e., employing some level of human effort. The most common practice is to utilize online crowdsourcing services to hire remotely-located gig workers (mechanical turks) to perform the annotation task on demand [55, 229]. This strategy has been adopted for both over-the-web collection and direct image capturing to adapt the data to the supervised learning needs.

5.2.1 Discussion

Online data collection and/or annotation involve two main risks against the playbook of data-centric AI. These risks are ground truth annotation inconsistency and inclusion of images/labels that may be heavily biased and violate individuals' privacy and rights. Even though image labeling may not sound like a fully subjective task, in practice, it can be carried out in different ways and levels of detail depending on the practitioner and the granularity of the curation instructions. As mentioned, labels inconsistency significantly lowers the potential of the data at hand. Moreover, recent studies revealed that publicly available visual data sets, commonly used in machine learning, contain verifiably non-consensual images of children, inappropriate and prejudicial adults material, as well as offensive classes that reflect obsolete views promoting discrimination and non-inclusivity [30]. Especially for large-scale datasets, these incidents may be amplified and difficult to detect, dropping the dataset's quality and possibly creating a chain reaction through the potential success and broad use of the collected set and/or other practices like transfer learning.

Most prominently, all the above reveal the culture that has been growing in the visual AI community. Until today, it is considered normal the appropriation of real people images as material for learning applications with no or minimal form of scrutiny. However, such datasets which are built under this (lack of) principle can directly or indirectly impact people when used in reality for training and validation. As Birhane and Prabhu point out, *"The absence of critical engagement with canonical datasets disproportionately negatively impacts women, racial and ethnic minorities, and vulnerable individuals and communities at the margins of society"* [29].

Of course, not every online data collection routine is problematic, and similar annotation, privacy, fairness, and bias issues can arise during standard image capturing procedures. However, it is beneficial for the community to start gaining awareness on where the large-scale real data for visual machine learning are coming from and what practices have been employed. Questions like these have previously hardly been considered. However, they are necessary to build proper curation processes. Ideally, data collectors should invest into curation and careful documentation rather than consuming everything on the web and would include pre-development checks on evaluating how the intended dataset fits into research and development goals and supports stakeholder values. To this end, image synthesis can be a step further. When well-defined within the data-centric practices, it can address the mentioned risks and provide increased levels of data coverage and diversity, scalability, and control over the definition and generation pipeline. Synthetic data can preserve (or can be tailored to preserve) privacy while they always come with automatically generated, perfect, and consistent ground truth annotations.

5.3 Data generation

It is clear that image data collection for supervised deep learning can be insufficient in several ways. The acquisition process can be cumbersome, requiring special sensors, infrastructure, and conditions. Capturing adequate samples of all possible situations that a visual dataset should cover is a strenuous process, and in some cases, even impossible, e.g., traffic accidents for autonomous driving. Most importantly, data collection means manual, or semi-manual, ground truth annotation. This is an expensive and labor-intensive operation that often prevents the construction of appropriate, in terms of quality and quantity, datasets. Issues like these drove the development of data-centric approaches that aim to maximize the potential of available data, like data augmentation. Although this category of methods is effective and has seen impressive growth through a wide range of solutions (section 5.4), they are bound by the data at hand. Data generation is an alternative that has become increasingly popular since the dawn of deep learning. This increase is justified partly by the fact that deep learning typically expects large training sets, and hence the need for data is more intense.

Apart from the clear benefit of computer-generated visual data on training algorithms, synthetic images can also be advantageous to evaluate models. It is now feasible to create high-quality images, with increased control over the content and generation process and the possibility to produce a large variety of meta-data. These properties enable thorough statistical testing and analysis that focuses on different dimensions. This means that the models under development can be tested on targeted content under various circumstances and domains that may be hard to occur or even out of the training data distribution. Image synthesis techniques can automatically associate to various forms of meta-information to the resulting data. These are related to, e.g., simulated sensor parameters, depth, and more, and can help on both data and model performance analysis and improvement. It should be noted that as for synthetic data for training, so for testing, and benchmarking, they must be representative of their real counterparts and the task being performed.

Naturally, the possibility of synthetic data triggered the research on how they should be formulated. To this end, there have been explorations of the benefits of oversampling hard cases or how unrealistic results could promote the optimization process. As a general guideline, synthetic data for training or testing deep learning algorithms should meet some principal characteristics. First and foremost, the generated images must be diverse. Training or testing will not be successful if the features in the produced imagery do not cover the target domain densely enough. Moreover, the resulting data would ideally close the domain gap between training and testing data domains. The minimization of domain shift can be incorporated in the synthesis pipeline or effectuated on the output images via adaptation methods. It is also desirable that the synthesis process scales easily. Given that many applications often require large amounts of samples, the generation pipeline needs to scale efficiently to accommodate the data requirements. Lastly, and of fundamental importance for supervised machine learning, is the supplement of high-quality, automatically generated labels. The benefit

of automatic, perfect, and consistent annotations is one of the main reasons why image synthesis is a central practice of data-centric AI development.

It has to be emphasized that synthetic data and thus this thesis do not intend to degrade the value and significance of captured images. Image synthesis aims to advance data-centric strategies, model analysis and ultimately empower successful application development and deployment. The end-user of visual deep learning applications is most often, if not always, reality. Synthetic data can simulate and possibly enhance the end domain, constituting a powerful tool to enable high-performing methods that can also facilitate technical or societal aspects, which in cases would be impossible to be satisfied otherwise. This thesis deals with such cases, described in the following sections, and proposes synthetic data solutions generated by computer graphics and generative modeling.

5.3.1 Computer graphics

Data generation has been employed increasingly in several computer vision tasks solved with deep learning. Computer graphics has been the means for this in an extensive range of applications (section 2.2), utilizing the basic concepts of modeling and rendering, as described in paper **A** and chapter 3.

Computer graphics-based visual data for deep learning applications build virtual content that may consist of geometries, materials, light sources, optics, sensors, and other features. These, depending on the application and approach, may range from simple setups [245] to fully-featured photorealistic virtual worlds, as shown in paper **B** and its successor released dataset [274]. Numerous data generation methods build complete world representations where the simulated sensors navigate and capture images, videos, or other measurements. This strategy is found typically in synthetic sets that come out of 3D development platforms [71, 203, 244]. These platforms usually provide geometry and material modeling tools, animation mechanisms, and scripting utilities. These functionalities mainly rely on manual efforts and artistic work. Although a pre-modeled virtual world can be considerably large containing dynamic objects, the resulting data diversity is finite. Procedural modeling engagement helps to overcome this bound and increase data variation. This way, the content is created on-demand while the final image can render only the geometry, materials, and lighting environments that are visible, directly or indirectly, to the virtual sensors to accelerate the per-frame rendering time.

Apart from the scene content alternatives, data synthesis has been associated with different rendering methods. The first one is based on real-time rendering, that employs mainly rasterization, and refers to either extracting frames or sequences from game environments [200, 201] or creating them directly on game engines and real-time simulators [95, 203]. Extracting data from existing commercial games requires software modifications that affect their rendering pipeline on the GPU. These are dedicated solutions that change or tweak one or more aspects of the game. Moreover, game engines' wide accessibility over the past few years transformed them to also a source of producing synthetic data. Today, game engines offer advanced modeling and dynamic interaction

mechanisms, special effects options, and rendering capabilities. These make them an appealing choice for generating data in high frame rates. Similarly, 3D development platforms and other types of simulators, like driving or physics engines, got attention as data generation methods [61, 173, 214, 219]. These might have been developed initially to serve different purposes but are suitable for synthetic visual data generation or collection. Modern versions have started to provide real-time ray-tracing capabilities based on sampling schemes with denoising algorithms.

Image data have been generated offline too, enabling physically-based ray- and path-tracing, apart from simpler light transport approximations like rasterization and ray casting [89, 90, 161, 245]. Offline renderers used for synthetic data generation are available either stand-alone or integrated to open-source or commercial 3D software suites.

Paper A identifies another popular method of generating training data that is not represented entirely by classical rendering techniques. This is *object infusion* and refers to training data synthesis methods that render single or multiple objects offline and infuse them on a background image to composite the final result [60, 232, 259]. Frameworks that apply a cut-and-paste fashion to image synthesis are also included in this family. In these, the resulting images are constructed by removing one or more objects from a source, possibly modifying the extracted components, and finally inserting them to the same or a new background [187].

Two principal concepts have proved to be efficient when designing and generating data for deep learning. Both of them are tightly related to bridging the domain gap between synthetic and real imagery, and between the training and testing sets. *Domain randomization*, as the name implies, randomizes the content in the simulated environments aiming to fill the reality gap between training and testing data [238]. The idea is applicable in principle in any data generation method that creates images from scratch. Similarly, *rendering randomization* randomizes the lighting conditions and camera configurations, i.e., the light sources' number, type, position, and intensity, and the camera extrinsic parameters and trajectories. The insight behind the modeling and rendering randomization is that the training data variability, provided by randomness, will make the real world appear to the model as yet another variation.

Among the various machine learning tasks, recognition has utilized data generation to the largest extent. This observation is can be partiallz attributed to the autonomous driving outbreak since 2016, where semantic segmentation [10, 71, 89, 90, 200, 203] and object detection [71, 119] gathered a lot of attention. Large-scale synthetic data that come with pixel-perfect automatic or semi-automatic annotations was the perfect solution for tasks like these, when annotation of real captured data has been reported to need 20 to 90 minutes per frame [40, 51]. Aside from recognition tasks, intrinsic decomposition [225, 266], depth estimation [10, 259], optical flow [41, 60, 161], and human body pose estimation [44, 65, 182, 223] have also used image synthesis extensively. Following the literature, non-procedural physically-based modeling is a common choice when it comes to building virtual worlds while real-time and offline rendering are both

frequently used synthesis options.

5.3.2 Generative adversarial networks

GANs for image synthesis is an emerging, powerful tool designed to generate new images out of some data distribution using a minimax strategy (chapter 4). This technology, though recent, has already demonstrated impressive, realistic synthetic images [39, 124, 126] but limited applicability as the pure source of training data in visual deep learning, compared to computer graphics (section 2.2).

Several GAN variants have been proposed for image synthesis. At present, the most successful applications of GANs are used in computer vision, including image translation [111, 129, 263, 282, 294], super-resolution [143, 264], matting [157], blending [275], inpainting [281], human face synthesis [38, 105], video generation [255], and more. In addition, they have been useful in other aspects of image generation, like texture synthesis [26, 146] and text-to-image generation [195, 288].

However, the use of GAN-generated image sets as means of training deep models for computer vision tasks has not yet matured beyond the domain adaptation and data augmentation practices (section 5.4). GANs may not require prior structural knowledge, as computer graphics methods do, but rely on a data distribution to learn to draw new samples from it. This means that the random noise-to-image GAN generation is tied, to some extent, to the underlying true data variation. Since the goal of training data synthesis usually is to expand diversity, the limitation is more apparent for natural images where innumerable feature combinations exist and less effectual in cases where the real data diversity is somehow bounded. This occurs in cases where the requested imagery needs to vary within a more defined domain with fewer features, like medical visual data of various types. Moreover, GAN data generation for supervised learning requires purpose-built configurations for advanced annotation types.

For these reasons, GAN synthetic data, generated directly from random noise, have not been widely adopted for supervised visual deep learning. However, some recognition applications have benefited, related to medical analysis [52] that usually use abstract conditions such as segmentation masks [221]. Among them, a few examples tackle face recognition [74, 247] and privacy-related issues [246].

One more dimension of data generation with the original GAN setting is the lack of control over content creation. This aspect can lead to problems with compliance with reality, 3D perspective and depth, and global structure coordination. For example, unnatural object/human shapes depicted in too flat or confused views are possible visual outcomes of GANs that are not always desirable or contribute to diversity. Advancements to increase control include modifying the GAN architecture by, e.g., disentangling the dimensions of the latent space [45] or controlling individual neurons of the generator that correspond to interpretable concepts [22].

Another promising alternative that lately has started to arise is what is called the *hybrid*

generation techniques. To this end, mixtures/combinations of generative modeling and classical computer graphics methods join forces to create images that benefit from both. For example, adaptable simulation systems have been proposed that can learn generative models of synthetic scenes from probabilistic grammars, and then obtain the final rendered images, with their corresponding ground truth annotations, via conventional graphics engines [122]. Somewhat in the reverse logic, a 3D scene can be modeled with computer graphics and represented with limited or without materials, textures, and lighting information. Then, a GAN variant can play the role of the renderer and transform the representation to its final form. In this context, the GAN module transforms/refines an image from one domain to another, a task that can be performed with image-to-image mapping, in a supervised [111, 262] or unsupervised way [294]. This type of transformation has been employed for realism enhancement purposes of computer-generated data [202, 224]. Lastly, image composition can be based on some available annotations. For this, a hybrid method can, for example, learn to construct segmentation masks and then assemble new images in an object infusion fashion [197].

5.3.3 Contributions

Based on the described data generation methods, this thesis contributes with a computer graphics synthesis approach for *data diversity*, exemplified for street scene semantic segmentation, and a generative image modeling method for *data privacy* that aims to synthesize anonymized digital pathology data.

Synthesis for data diversity

Paper **B** addresses semantic segmentation for autonomous driving and robotic navigation by introducing a systematic data generation method of highly realistic and diverse synthetic urban scenes. Semantic segmentation is crucial for visual scene understanding, and deep learning solutions typically require large-scale datasets with pixel-perfect ground truth annotations. Capturing large amounts of street-level views is possible, with the few notable examples to constitute the research standard [51, 175, 284]. Data collection for this purpose requires well-defined guidelines related to content variability and optics configuration, and above all high ground truth label granularity. Fine-grained annotations of big datasets for semantic segmentation is a strenuous and expensive procedure, where whether it is fully or semi-manual, carries the risk of errors and label inconsistency. In pixel-level tasks like this, the annotation quality is vital for performance and accuracy, and therefore coarse labels can negatively affect the end application. Moreover, it is often difficult to cover all cases, especially the hard ones.

These reasons have led to the growth and use of synthetic imagery for tasks like this. Data generation arrives with automatic pixel-perfect annotations while a rich variety of events and conditions can be simulated. The majority of synthetic datasets for automotive applications come from computer games, where real-time rendering of finite virtual worlds is employed, and some degree of manual ground truth annotation is

required. The synthesis/extraction method affects data realism and variability. Paper **B** introduced an alternative data production strategy that builds upon procedural physically-based modeling and physically-based rendering, enabled through path-tracing, to deliver realistic images that essentially represent a new simulated world per frame.

The proposed framework originated from the visual effects and film industry and proposed automatic world generation, via procedural modeling, with accurate light transport simulation in a scalable, cloud-based computation capable of generating hundreds of thousands or millions of images. Previous game-based methods used concrete virtual worlds to provide data for deep learning. The nature of computer games is exactly this, to use a static environment where some dynamic objects are moving around, and a few features are parameterizable. In the context of data generation for deep learning, all the constructing components would ideally be varied such that each image would be as different as possible from the rest. However, in the context of a computer game, the resulting image variability is tied to a fixed amount of possible combinations. To this end, paper **B** contributes with a procedural modeling solution that instantiates a unique world for each output image from a set of classes that represent vehicles, buildings, pedestrians, road surfaces, vegetation, trees, and other relevant attributes. All the pertinent scene factors, i.e., geometry, materials, color, placement, orientation, etc., were parameterized, and the synthesized image and its corresponding annotations are a sampling of that parameter space. This configuration provided advanced control over the variation of the classes and relevant factors while yielding an exponential explosion of possible combinations/worlds, as each added factor of variation multiplies rather than adds to the parameter space. As a result, the proposed system produced a highly content-diverse dataset. The scene variation was enhanced further by changing the lighting conditions to include all times of the day. For the non-procedurally generated scene content, 3D models were fetched from a model library. The fetched models were modified locally, in terms of material and position, to induce more variety. The analysis has focused on the most central street scene classes for automotive applications: road, sidewalk, traffic light and sign, person, rider, car, and bicycle. These were designed with heightened variation, geometric accuracy, and material representation.

To produce a realistic image set, path-tracing rendering with Monte Carlo integration was engaged. However, this was not the only measure taken towards this direction. Realism injected into synthetic data is not a one-dimensional quality. It requires that the generation process allows realism in all separate steps of scene definition, geometry shaping, illumination and light transport simulation, object materials and textures, and optical effects. In the presented work, physically-based procedural modeling and rendering, combined with point spread functions in the image domain, handled each one of these aspects. Although it is more time-consuming than real-time rendering and iterating over the same geometries for multiple images, the system created and rendered only what was visible from each image's camera view, directly or through reflections, and cast shadows.

(a) Dilated fully convolutional network [283].

(b) Full-resolution residual network [188]. Note that for the proposed method, there is no representation for the classes *Wall*, *Fence*, and *Train* and, therefore, there is no reported performance. In contrast, in SYNTHIA and GTA all classes are present and, therefore, zero IoU means failure to recognize the respective class.

Figure 5.2: Per-class (bars) and overall mean performance (solid lines) for two network architectures. (a) employs ImageNet weight initialization, while (b) is training from scratch. The networks were trained with synthetic data — SYNTHIA, GTA and the proposed method — and tested on real data (Cityscapes). The corresponding performance using only real data (Cityscapes) is given for reference.

The resulting synthetic dataset was assessed with two state-of-the-art, at the time, networks [188, 283]. The evaluation experiments were set up to cover both the training modes of using only synthetic data for training (Figure 5.2) and finetuning on the real dataset of Cityscapes [51] (Figure 5.3). Moreover the synthetic data only experiments cover the cases of ImageNet weight initialization [283] (Figure 5.2a) and training from scratch [188] (Figure 5.2b). For this paper, the introduced framework synthesized 25,000

(a) Dilated fully convolutional network [283].

(b) Full-resolution residual network [188].

Figure 5.3: Per-class (bars) and overall mean performance (solid lines) for two network architectures that both employ finetuning from synthetic data — SYNTHIA, GTA and the proposed method — to real data (Cityscapes). The networks were tested on real data (Cityscapes) and the corresponding performance using only real data (Cityscapes) is given for reference.

images to be compared to two other popular synthetic datasets, denoted in this text as SYNTHIA [203] and GTA [200] consisting of 9,400 and 20,078 images respectively.

The most striking findings of the analysis are summarized in three points. Firstly, looking at the overall synthetic datasets' performance when they are the only source of training data (solid lines in Figure 5.2), it is clear that all three options perform poorly on the test set of Cityscapes. This outcome was expected as the scenario of training with synthetic datasets only is harder compared to the finetuning case where the weights initialization can smooth the gap between real and synthetic domains to some extent.

However, training with synthetic data alone reveals the true potential of each method. Looking at the datasets' in-between comparison, the proposed approach outperforms by far the rest of synthetic sets for both the architectures, with or without weight initialization. It is then inferred that both SYNTHIA and GTA exhibit a higher domain shift from Cityscapes than the presented strategy which performance indicates reduced domain shift and less dependence on ImageNet weights. Secondly, observing the per class performances when training with synthetic data only, the evaluation shows that the higher realism and feature variation applied in the selected classes of focus improves their semantic segmentation accuracy. Specifically, seven out of eight of the central street scene categories reach the highest score in both architectures. It is beneficial to assess the high-performing classes in conjunction with the low-score ones (zero performance) reported for SYNTHIA and GTA. It is derived that adequate exemplars representation is required for the network to learn while refining their constructing elements boosts performance considerably. For the latter, the procedural modeling approach ensures adding more variation where needed. Lastly, it is apparent that when finetuning on the full set of Cityscapes its features become dominant, as expected. However, there are benefits from pretraining on synthetic data, especially for the proposed dataset and for the classes with a low representation in the real set, like traffic lights, walls, fences, and trains. Most notably, training with the proposed dataset, which does not contain any samples of the wall, fence, and train categories succeeds in improving these classes' accuracy significantly (Figures 5.2b and 5.3b). This is explained as the networks share weights across classes and, thus, when a wide range of data is seen during training, several classes can benefit from improvements to the shared set of features. This way, finetuning on real data from a refined synthetic set heightens performance for (any) targeted categories and weaker classes.

Synthesis for data privacy

Paper F focuses on the problem of synthesizing training data using GANs, with the purpose to be the sole data source for deep learning downstream recognition tasks. This setup is driven by the privacy-preserving and anonymization condition that some applications need to fulfill. Examples refer to applications that utilize sensitive or protected data, like medical image information which cannot be fully released, and need to ensure the subjects' private data protection. The medical imaging case, focused on digital pathology in this thesis, except privacy protection that may be required, most commonly suffers from training data deficiency. It is, therefore, beneficial to develop methods capable of compensating for the shortage of supervised training data of this type, that at the same time, would be in line with any privacy requirements.

For this purpose, paper F uses generative techniques and studies the impact of using ensembles of GANs to synthesize complete standalone training sets for deep classifiers. This formulation diverges from the typical practice that employs both real and synthetic images in some joint scheme where usually the computer-generated imagery is perceived as complementary variations to the substance of captured data features. In the case of GAN-generated training data, the synthetic images need to enclose all the

necessary information themselves, making diversity a critical aspect and mode drop an important problem to be addressed.

While some GAN variants have been introduced to improve mode coverage, another propitious solution is to train ensembles of GANs with boosting strategies that force them to focus on different parts of the data distribution [84, 240, 265]. Paper F builds upon this idea and shows that to synthesize training data for deep models, the most influential aspect is to combine multiple GANs trained *independently* without enforcing any sophisticated boosting schemes. It proves that even though some of these boosting schemes are in principal intuitive, in practice, they can result in similar or slightly worse performance and are more prone to overfitting.

The observations came from three sets of experiments that tested the ensemble performance on 2D data, two low-resolution natural image sets, and the most realistic scenario of anonymization of breast cancer histopathology images. The example of 2D data is a simple yet effective way to provide insights for the mode coverage capabilities, while the rest sets of experiments allow for ensemble performance analysis under different GAN complexities. For this, deep convolutional GAN (DCGAN) and progressive growing GAN (PGGAN) were used in different ensemble sizes, a factor that proved of importance for the classification performance, with increasingly larger ensembles to perform better. For the most practical case of privacy-preserving medical data synthesis, the ensembles of independently trained GANs performed on par with the boosting strategy. This reveals that this method may be less complicated but effective in improving the generated quality and coverage variation. One benefit of this approach over boosting is the low risk of concentrating on a fixed part of the data distribution. This could be potentially a problem with boosted GANs, leading to overfitting on the exact images used for training, a situation that oversteps the objective of anonymization. Even though the empirical evidence highlighted the importance of ensembles of GANs to generate synthetic training data capable of being used in isolation from their real counterparts, there is still a significant performance gap with captured data. Thus, future work on methods that improve GAN-based image synthesis for training deep models is needed.

5.3.4 Discussion

Diversity in synthesized data is one of the most important aspects of data generation, especially when they are going to supply deep learning models. Simulating natural imagery with computer graphics to create visual results that appear like captured requires a fair amount of the cases to model high-dimensional feature spaces and simulate effects inherent to the sensor and the capturing process. The presented approach for street-level semantic segmentation demonstrated that *realism* is a factor of significance for this type of data and applications. The evaluation analysis showed that highly varied and realistic generated images, both in terms of low-level (antialiasing and motion blur) and higher-level features (accurate geometric modeling and light transport) significantly improve the performance. This is clear when the dataset development is monitored on a pure synthetic training data basis, and the relative merits compared to less realistic

alternatives can be assessed down to the per-class performance. In this context, the introduced procedural scene modeling engine coupled with physically-based rendering has the potential to maximize both the diversity and realism of synthetic data.

On the other hand, it is natural to approach training data image synthesis for medical purposes from another perspective. This type of data, in particular histopathology images that this thesis explores, has a narrower domain of variation compared to natural images, while realism, in the sense that it was defined above, is not relevant. The presented work using generative modeling revealed that ensembles of independently trained GANs can perform on par, if not slightly better, with boosting methods, suggesting that this method, even though simpler, can be a viable solution for privacy-protecting data of this kind.

There are several possible future research directions for both image synthesis methods for training data purposes and applications separately. For example, related to paper **B** that deals with natural, computer graphics generated images, it would be interesting to analyze more in detail the impact of realism — in terms of light transport simulation — on the network’s performance to understand the tradeoff between computational cost and inference results. Regarding paper **F** that proposes image synthesis of anonymized medical data using ensembles of GANs, it could be further explored the impact of the training data size, while there is also an obvious need for establishing standardized quality measures tailored to this specific application of generative models, and more comprehensive evaluations of existing techniques for diversifying GANs in the context of synthetic training data generation.

Finally, it is worth mentioning the potential of the hybrid synthesis methods that open up ahead. These leverage both computer graphics and generative modeling techniques and can, in principle, facilitate several image modalities including natural and histopathology visual data forms. For example, a hybrid method for street scene parsing could build on the proposed procedural engine for scene modeling and per-pixel accurate ground truth annotations and then rely on some GAN variant to produce the final results. Similarly, there are hybrid methods for histopathology image generation and nuclei segmentation in specific, where the suggested pipeline employs both GANs and real textures patches. An initial synthetic image patch is generated through classical processing methods, and then it gets refined through a GAN module [104].

5.4 Data augmentation

While model-centric approaches focus on the model’s architecture to increase generalization — by progressively building more and more complex networks and sophisticated regularization techniques — another effective strategy is to address the problem at the root, i.e., train the model on more, quality, and relevant data. In practice, the available visual data for a specific application are usually limited, and in some cases, such as medical image analysis, very limited. Data augmentation is a data-centric solution to

increase the training set's size, variation and diversity by constructing fake/synthetic samples, most often out of the available images.

Image data augmentation is a popular and well-studied technique, especially for classification, recognition, and detection [68, 168, 184, 220, 222]. The reason is straightforward. A classifier has to map an n -dimensional input x to a single label y and, therefore, needs to learn to be invariant to a wide range of transformations. Images are high dimensional, and as such, they provide a large area of variation factors, many of which can be simulated. Data augmentation generates new training pairs (x', y) by performing simple or more drastic transformations on the available training samples. The new pairs inflate the training set either by data *warping* or *oversampling*.

Data warping methods apply transformations to the existing training set while preserving their labels. These include classic image manipulations, like geometric and color transformations, and other types like adversarial examples, and neural style transfer. On the other hand, data oversampling augmentation methods *synthesize images* to complement the training set. A typical example is image synthesis with GANs. Warping and oversampling augmentations do not exclude each other. On the contrary, it is common to stack various kinds of augmentation to achieve higher levels of variation, e.g., a complete data augmentation strategy for an imbalanced training set could be adaptive duplication stacked with geometric (rotation, translation, etc.) and color transformations, and combined with noise injection (Figure 5.4).

Generally, not all image augmentation types are label-preserving. When an augmentation scheme is incorporated into the learning pipeline, it should be applied according to the downstream task and in ways and amounts that ensure label preservation or adjustment. Considering, for example, image classification, where some forms of the aforementioned geometric transformations are entirely, or under simple terms, label-preserving, like flipping, whereas others can lead to image-label mismatches, like spatial shifting operations and color transformations that can severely affect the subject of interest.

In the following sections, the various augmentation methods, applied in the image space and commonly used in deep learning training procedures, are described in detail. Of these, feature space augmentations that manipulate the internal/intermediate representations of a neural network are excluded [222].

5.4.1 Image manipulations

Currently, the most popular data augmentation style is combinations of affine image transformations and color alterations. Kernel filtering, partial erasing, and image mixing constitute other operations that manipulate the training data. All of them are conceptually and implementation-wise straightforward, fast, and reproducible with a positive effect on the generalization error [184, 237]. Typical image manipulation augmentations include geometric and photometric transformations, noise injection, and cutout and mix-up strategies.

Figure 5.4: Image manipulations as training data augmentations. These types of image processing, like **(b)** geometric transformations (rotation, translation, flipping, etc.), **(c)** color jittering, and **(d)** noise injection, usually stacked in combinations **(e)**, are considered today a standard practice, applied nearly to any visual deep classifier.

Geometric transformations

Image *translation* by a few pixels in each direction has proven to be a helpful transformation against positional bias, even in the case of CNNs, where the deployed model is already partially translation-invariant due to the convolution and pooling operations. Other functions, such as *rotation* in various angles, *flipping* horizontally and vertically, center or random *cropping*, *shearing* and *scaling* can also be quite effective. If any image space remains empty (without content) after the transformation, it is common to fill it by either a constant or the nearest color value, or by mirroring or wrapping local patches or noise. Thus, the spatial image dimensions stay the same. Lastly, *elastic deformations* can also increase the performance of classifiers, in addition to the aforementioned affine transformations [123, 226]. These are stretches or contractions of small regions in the image performed by a normalized random displacement field.

Photometric transformations

Lighting bias is common to image recognition problems. Photometric augmentations increase image variation by manipulating the color and lighting while keeping the geometry intact. *Color jittering* uses random color manipulation [276] or fixed color adjustment [194, 237]. Color changes can be done in a channel independent manner or at color space-level, e.g., select an appropriate color model (RGB, HSV, etc.) and modify all channels according to a defined rule [230, 231]. Moreover, *fancy PCA* is a color augmentation scheme that performs principal component analysis (PCA) on sets of pixels throughout the training set to find the principal components and relative color balance. Then the training images' color pixel values are shifted by adding multiples of the principle components [135, 237].

Kernel filtering is another strategy inspired by image processing. Compared to the rest of the image manipulation routines, this one is underexplored in deep neural networks, possibly due to its similarity with the internal CNN's mechanisms and potential to be part of a network layer instead of data preprocessing. However, typical examples are

blurring and edge enhancement filters [237], while more sophisticated solutions that act both on image and feature space have also been proposed [121].

Noise injection

Injecting the input data with random *noise* can improve neural network robustness. The procedure is simple: random perturbations are added to the color channels of each pixel in the image. The Gaussian noise model is most frequently used [220], but other types have also been explored [18].

Cutout and mix-up

Cutout or *random erasing* modifies the training data by randomly selecting rectangles in the input images and erasing/replacing the cutout pixels with random values (noise) [56, 293]. The result is training image variations with different levels of occlusion, reducing the risk of overfitting while at the same time improving occlusion robustness. This method can be perceived as a form of dropout on the image level. The difference is that a spatial prior is applied in the input space, and the operation is performed on a continuous region.

Another strategy is *mix-up* which is based simply on combining pairs of images by either overlaying them and taking the average for each pixel or slicing parts and concatenating them in various schemes (vertically, horizontally, randomly) [110, 234, 239, 289]. The concept has also been extended to include four images, where they are randomly cropped and mixed to formulate a new training sample [236]. Regarding the class labels of the augmentation images, these methods also mix the annotations resulting in soft label versions.

Finally, there have been methods proposed that belong to both mentioned categories combining both principles, i.e., patches of the available images are cut and pasted among the training set, while the ground truth labels are also mixed proportionally to the area of the patches [286].

5.4.2 Deep learning approaches

Apart from image processing methods that can be utilized for augmentation purposes, deep learning itself can be used to modify or learn to modify, usually by some form of mapping, the available data. The result is new synthetic samples suitable for training data augmentation purposes. Such deep learning methods are adversarial training, neural style transfer, and GANs, which are part of this section.

However, they could not be left without mentioning other learning and meta-learning augmentation practices that, for example, locate the most fitting augmentation approach(es) for a defined task, given a search space of different transformations [54]. For further details, Shorten and Khoshgoftaar provide a broad review [222].

Adversarial training

Adversarial attacks are the art of fooling deep neural networks developed for typical computer vision tasks, e.g., recognition. Goodfellow et al. [80] showed how to trick a trained neural network to misclassifications by minimally changing images. Their method, known as the fast gradient sign method (FGSM), computes the loss gradient with respect to the image pixels and uses it to modify only the most dominant pixels. This results in adversarial images/examples that are almost impossible for humans to distinguish from the clean images but are classified wrongly by the net. The concept was extended to unlabelled samples, too [170]. Other methods range from one-pixel [233] to universal adversarial perturbations [172]. For further details on attack and defender methods, Akhtar and Mian provide a comprehensive review [6].

The adversarial images can complement the training data to improve robustness against perturbations and domain shift and adaptation [260, 277, 292], in some cases at the cost of standard accuracy on clean images [248]. Even though *adversarial augmentations* can be constrained from small changes that keep images visually unaltered, to modifications that would not easily occur to a test set, they aim to improve weak spots in the learned decision boundary. Some examples are AugMix [93] that randomly mixes multiple diverse augmentations, DeepAugment [94] and its successor [43] that optimize the parameters of image-to-image models to generate diverse images, and MaxUp [77] that generates hard samples by selecting the worst-case weights from the mix-up image manipulation strategy [289]. The latter is designed for both generalization strength and adversarial robustness, and it has recently been combined with AugMix to unify diversity and hardness [261].

Neural style transfer

This is an optimization method that takes one content and one style image and creates a new one that depicts the content image modified with the style of the corresponding reference image. The result is achieved by optimizing the generated image to match the content statistics of the content image and the style statistics of the style reference image. The statistics are extracted using a CNN. Several advances, including the use of perceptual loss and instance normalization, enabled *fast stylization* [72, 106, 118, 256]. For a general overview of the neural style transfer algorithms, loss functions, evaluation methods, and comparisons see the review given by Jing et al. [117].

For data augmentation, among the recently proposed methods, it is worth mentioning Style Augmentation that uses style transfer to augment arbitrary training images, randomizing their style, i.e., color, texture, and contrast, while preserving geometry. Deriving reference style at random from a wide distribution helps to avoid introducing a style bias to the dataset while proving to have a positive impact on test performance, particularly in the presence of domain shift [112]. StyleMix is a combination of mix-up image pair manipulation with separate controls for the content and style components creating a version of adversarial augmentations that provide improved model robustness [102]. Neural style transfer augmentations' applicability range from computer

vision applications of natural images [102, 112] to various kinds of pathology and medical image modalities [49, 168].

GANs

Without a doubt, there is an ongoing, animated development of GANs where it is now possible to synthetically generate high-resolution images, almost visually indistinguishable from real captures. In the past five years, they have flourished as an augmentation technique, too, to boost the performance of low-data scenarios and compensate for imbalanced training sets for image classification. Even though there are applications of natural images that utilize GAN augmentations [15, 151, 158, 184], particularly oriented to human faces [9, 13, 295], the largest body of research concerns the medical imaging and digital pathology domains [36, 70, 85, 88, 213, 221, 279, 280]. There are two interconnected reasons for this. The collection and curation of medical visual data suffer from several dependencies, such as capturing infrastructure demands, large amounts of annotation time for pathologists, and patients' privacy. Thus, the available data are often limited, increasing the need for techniques to enhance the collected data distribution. At the same time, generating histopathology images with GANs exhibit less complexity than natural scenes since the range of the features that need to be learned is narrower. This behavior is more pronounced in the case of noise-to-image generation compared to adaptation synthesis examples.

Common GAN flavors used for data augmentation are progressive growing GAN (PGGAN) [36], deep convolutional GAN (DCGAN) [70], and mostly conditional GAN (CGAN) [151, 221, 279, 280], and cycle-consistent GAN (CycleGAN) [85, 184, 213, 295], while there have also been proposed variations/modifications of the above specially designed to address problems like face recognition [9], semantic segmentation [15], or to act as generic methods for image classification [13, 158]. Figure 5.5 shows example CycleGAN augmentations for both natural and histopathology images.

5.4.3 Contributions

It has been made clear that data-centric AI can benefit from a wide range of training data augmentation methods. This thesis presents contributions to two very common problems. It deals with *data scarcity in digital pathology* and *class imbalance for natural images* using CycleGANs as the primary augmentation tool.

Augmentations for data scarcity

Papers **C** and **D** investigate the problem of metastasis detection of colon adenocarcinoma in the regional lymph nodes using deep learning-based classification. Along with grading the primary tumor, the assessment of the tumor spread to the lymphatic system is an important prognostic factor. Constructing large, annotated datasets for deep learning metastasis detection is particularly challenging due to the low-tumor-to-non-tumor ratio in the tissue. This means that it is likely that many areas in a WSI will be

Figure 5.5: Deep learning augmentations, GAN-based in particular **(b)**, for the two image modalities studied in this thesis. For these, the level of difficulty and success of GAN-generated samples, generally, is not the same. CycleGANs can be used to map, for example, complex natural images from one weather condition to other modes **(top)**, and breast cancer histopathology images to other organs' cancer representations **(bottom)**. This way, new synthetic samples can enhance the training data distribution.

of healthy tissue, and only small parts may contain malignant cells. Therefore, looking for cancer metastasis in the lymph nodes requires exhaustive monitoring, making the annotation process a strenuous and time-intensive procedure for pathologists.

Papers **C** and **D** study how data with lower acquisition and annotation cost can be used to augment very scarce training sets, in some cases under the absolute absence of target domain samples. For this, two augmentation strategies are proposed. The first one is to use inter-organ augmentations, i.e., annotated data from different organs from existing public datasets. The second one concerns intra-organ augmentations using the primary colon cancer tumor that is surgically removed and known in advance for malignancy. In primary tumor the tumor-to-non-tumor cells concentration is high, and the annotation process is faster.

To derive safe conclusions on the effect of the proposed augmentations, we set a data-centric controlled environment. This consisted of a simple deep classifier with some standard geometric and color jittering augmentations, that were kept fixed, while the training set that was fed to the network was altered. The alterations were based on three different data availability scenarios, defined by the annotation cost (measured in time

and effort). Intra- and inter-organ schemes were implemented in two ways. Either as a direct mix to the available data or as a GAN-based augmentation where the out-of-target domain data was transformed to the target domain via image-to-image translation and CycleGANs. In particular, this means that the primary tumor and the tissue from other organs, breast and skin, were added to the available colon data either as they were or transformed to lymph node tumor using CycleGANs. Breast and skin were selected to demonstrate different levels of model-perceived similarity to the target data, a factor that proved to be of importance for CycleGAN domain transformations. In this context, the similarity is related to the representation shift between the source (primary or other organs) and target (colon) data, calculated as the distance between the learned representations of the model. Moreover, two size schemes for the augmentation sets were tested in each respective scenario and defined experiment, leading to 985 trained classifiers and 52 CycleGANs in total.

Figure 5.6 summarizes the main findings of papers **C** and **D** that can be verbalized in three points. The first one is that primary tumor is an inexpensive, valuable training data source for metastasis detection. Even with no target domain representation (lymph node colon cancer), it is possible to detect it. Furthermore, a training set of primary tumor samples augmented with breast lymph node tissue is equivalent in performance with a larger set consisting of only target data, and even improved in robustness. This paves the way for similar possibilities in other cancer types using the TNM staging protocol [37], such as breast and esophagus.

The second insight is that out-of-domain data improve robustness, as measured by the performance's standard deviation over multiple trainings. Primary tumor and other organs' tissue provide a boost in performance, additionally to classical augmentations. Meaning, out-of-domain data have a regularizing effect on the optimization process leading to a significant impact on robustness.

Lastly, domain adaptation closes the gap between the target domain and the augmentation alternatives with a large representation shift. CycleGAN-based transformation of tissues with already low representation shift compared to the target distribution, e.g., breast tissue, does not improve performance or robustness. In this case, direct mix augmentation increases the data diversity sufficiently to outperform the baselines. On the other hand, skin data, exhibiting a much larger representation shift from lymph node colon cancer, show a significant avail from data adaptation. In this case, direct mix drastically decreases performance, indicating that the added diversity, as is, does not contribute to convergence. It is then better not to perform data adaptation in cases that will decrease the regularization effect of the out-of-target domain data (e.g., primary tumor or breast data with low representation shift). However, if the data are widely different (e.g., skin data, with large representation shift), it is necessary to perform adaptation to benefit from augmentation.

Figure 5.6: Performance, measured as mean accuracy under the receiver operating characteristic curve (AUROC), against a robustness proxy defined as the AUROC’s standard deviation over multiple trainings. The main observations suggest that 1) primary tumor is an inexpensive, valuable training data source, even in the absolute absence of the lymph node colon cancer target domain, 2) the out-of-domain data improve both performance and robustness, even when added in their raw form (especially lymph node breast tissue), and 3) CycleGAN domain transformation is particularly beneficial for out-of-domain data that exhibit a large representation shift from the target domain (like skin tissue).

Augmentations for class balancing

Paper E extends the concept of inter-domain transformations to natural images. Specifically, it is a systematic evaluation of GAN augmentations to complement small, imbalanced training sets for single-image multi-class weather classification. Conventional augmentations is a critical component in cases of imbalanced training. However, their effect on performance is limited. GANs have proved to be a successful augmentation tool, mostly for data of limited diversity, such as in histopathology images, as showcased above. This paper investigates the possibilities of generative augmentations to balance a small weather classification dataset, where one class is imbalanced, i.e., has a reduced number of images compared to the rest. Intra-class augmentations employing classical geometric and photometric transformations as well as noise-to-image GANs are compared to inter-class augmentations where images from another class are transformed to the underrepresented class using CycleGANs.

This investigation used three different GAN flavors. PGGANs, CycleGANs, and an extended version of CycleGANs that integrated weight demodulation inspired by Karras et al. [126]. The latter is suggested to likely improve artifacts in the generated images, identified as noise or the droplet effect. PGGANs were trained on the available samples of the underrepresented class and used to exemplify intra-class GAN augmentations. Even though they are found in the literature as an image synthesis method for training data augmentation in medical applications [36], for the more complex case of natural images, it turned out to be more challenging to produce data of sufficient quality for training. This was particularly profound in cases where the imbalanced class was very scarce. As an alternative, CycleGANs were used to introduce inter-class transfer. The approach of mapping images from a fully represented class to an underrepresented one resulted in increased generated diversity compared to the noise-to-image image synthesis. The reason is that the synthesis process was not limited to a narrow domain of a few images of one class. Moreover, CycleGAN transformation of an already existing image is a more trivial problem than PGGAN image generation from scratch.

The showcase application is a supervised deep classifier that determines the weather condition in an image out of three (sunny, foggy, snowy) or four (sunny, foggy, snowy, rainy) classes. In these experiments, different levels of imbalance for two classes (foggy and snowy) were studied. The GAN-based intra- and inter-class augmentations were compared to classical (intra-domain) methods consisting of common image manipulations and the simplistic strategy of image duplication. The experimentation of 1960 classifiers and 18 GAN models is summarized in Figures 5.7a and 5.7b, for the four and three classes classification experiments respectively. All augmentation methods improve the overall performance significantly for all scenarios and especially for the extreme imbalanced cases where the available amount of images in the underrepresented class is equal to the 25% of a balanced one. Given this, the impact of the different GAN-based augmentations is most evident looking at the individual classes and particularly the underrepresented ones. There are three main observations. The first one is that the best performing method is CycleGAN that achieves high accuracy levels for the underrepresented classes even for the extreme imbalance scenarios. Secondly, PG-

(a) Four classes weather classification.

(b) Three classes weather classification.

Figure 5.7: Augmentations to balance the class *foggy* (top row of (a) and (b)) or *snowy* (bottom row of (a) and (b)) in weather classification experiments with *four classes* (a) and *three classes* (b). Three imbalance levels were explored per case (equivalent to 25%, 50%, and 75% of a balanced set) and tested both *intra-class*, i.e., duplicates, geometric, photometric, noise injection, and PGGAN, as well as *inter-class* augmentations, i.e., CycleGAN and its variant that incorporates weight demodulation (WD).

GAN is overall the worst-performing GAN-based augmentation strategy. The reason is that the generated images come from the noise-to-image mapping where only the under-represented set is available. This leads to limited feature variation compared to CycleGANs that draw samples from another domain (another class) and transform them to the underrepresented class. Regarding the CycleGAN variation that integrates weight demodulation, there is no concluding benefit in performance compared to the vanilla flavor. Finally, and on a general note, the inter-class GAN augmentations are consistently better than the intra-class ones, most prominently for the heavily imbalanced scenarios.

5.4.4 Discussion

Data scarcity and class imbalance are most often two interconnected problems. Data augmentation is a straightforward and efficient solution whose execution can vary and adapt depending on the case and the image type. The presented augmentation approach for low-data in histopathology may be specific to the colon adenocarcinoma metastasis detection, however, the problem of training data scarcity is universal in digital pathology. Gathering and curating lymph node data for deep learning is a cumbersome process. The presented work aims to open up the road to further research on the proposed intra- and inter-organ augmentation schemes for metastasis detection of other cancer types, too. This aspiration counts on the uniformity across staining, scanning, and annotation protocols of the already existing, available histopathology cancer data and the ability to enhance them with further regularization methods to leverage their full potential.

To this end, it is important to notice that using inter-organ tissue augmentations requires adjustable strategies depending on the selected domains and their resemblance to the target data. For example, CycleGAN adaptation of data that are “*closer*” to the target distribution, according to some similarity metric, may offer nothing over direct mix and have an even detrimental effect instead. Conversely, for tissue data that are further away, domain adaptation is necessary. These observations call for future work to define appropriate frameworks and guidelines on when to adapt and when not. These would need to determine unified quantitative measures to compare model-specific differences between datasets, such as the representation shift used in the presented experiments. Moreover, other data types, augmentation methods, model-specific adaptation approaches, and transfer-learning techniques are yet to be explored. Based on the presented contributions, it is natural to predict that inter-organ formulations have the potential to be a valuable tool in future machine learning-based medical diagnostics.

While GAN-based augmentations have been used broadly in medical imaging applications, data scarcity and class imbalance in natural image sets have been treated mainly through classical image manipulations so far. For the most part, the reason is that GAN synthesis, and especially noise-to-image generation, is a complex task for the high-dimensional feature space of natural images. The presented work in single-image multi-class weather classification shows that GAN-based methods, inter-class CycleGAN transformations in particular, can be more effective than the standard augmentation schemes

(geometric, photometric, noise variations). This paves the way for further GAN-based augmentations in scenarios where the data is both diverse and scarce.

This potential reveals several venues of further work related to GAN augmentations for class balancing in small sets of natural images. For example, one direction would be to formulate image selection protocols for the augmentation step based on the GAN outcome image quality or use this metric to weight the loss function in classification. Moreover, it would be interesting to investigate if sophisticated methods that can improve image formulation and extend diversity could be incorporated in the generation process of a CycleGAN. Finally, more combinations could be tested to define the best possible scheme, which can be, for example, a stack of intra- and inter-class augmentation strategies.

It is safe to conclude that the concept of inter-domain transformations is a promising strategy for scarce and/or imbalanced sets of both histopathology and content-rich natural images. GAN-based image synthesis, with a special focus on CycleGANs that can leverage already existing data, has the potential to become a powerful tool in the data-centric pipeline of supervised visual deep learning that deals with low-data scenarios.

6 Conclusion

This thesis, together with the work presented in the accompanying papers, provides a structured and broad view of the data component in the visual machine learning equation (Figure 1.1). Several aspects of the typical data-centric pipeline of supervised deep learning were described and discussed, from technical issues to potential ethical and societal risks. The core idea is using *synthetic image data* in the learning process, and thus special emphasis was given to the relevant areas of contribution. Synthetic image generation is approached from two perspectives, utilizing classical computer graphics techniques, that construct from scratch 3D virtual worlds through modeling and rendering methods, and generative modeling, that learn visual representations out of given data distributions and can synthesize new samples through statistical and probabilistic methods.

Initiating from general descriptions of visual data, image synthesis, and deep learning, synthetic data were framed and motivated as a powerful part of the learning pipeline. The necessary background information on deep networks and the relevant application areas were briefed to set the proper foundation, conceptually and implementation-wise. These areas concern computer vision tasks, such as semantic segmentation and classification, for natural and digital pathology (histopathology) images. Subsequently, the two explored image generation methods were analyzed in more detail, focusing on the components employed by the conducted research. Procedural modeling and physically-based rendering were central concepts in the presented data generation system while three GAN variants were utilized for noise-to-image and image-to-image synthesis. Finally, all the different parts of the theory were tied together under the data-centric AI realm, where synthetic visual data were exemplified as the only source of training data or as the means to augment small training sets of real/captured images. The thesis contributions were grouped thematically under these practices intending to provide comparative and collective analysis and solutions, and highlight the underlying relevance between the individual papers. This approach enabled discussions in the light of the general context of data generation and augmentations, including limitations and possible future directions.

This chapter concludes the body of the presented work with a synopsis of the thesis' contributions and a general discussion around image synthesis.

6.1 Contributions

This dissertation deals with the idea of synthetic data for visual machine learning. In the introduction, a threefold objective was defined related to the explored image synthesis methods, image modalities, and the applied training modes. Recognizing that the latter provide different practices in a data-centric deep learning pipeline (Figure 5.1), this section positions the contributions in connection with the objectives under the training data generation and data augmentation modes.

6.1.1 Data generation

The majority of this thesis contributions in the field of training data generation, as standalone sets, concern natural images and computer graphics aspects. Generative modeling-based data generation is also explored in the form of ensembles of independently trained GANs and in comparison to boosting strategies for the scenario of creating histopathology data (and low-resolution natural image samples). In this mode, the contributions focus on data diversity which is one of the essential factors for successful models that generalize well in various relevant environments. Moreover, synthetic training data exhibit interesting and sometimes necessary privacy preserving properties as well. Specifically,

1. Paper **A** provides an in-depth overview of synthetic data generation methods based on computer graphic techniques for training visual machine learning algorithms. It contributes with a computer graphics taxonomy and an additional categorization according to the targeted computer vision tasks. The majority of the documented methods/datasets concern natural images while fewer random configurations of natural concepts and medical imaging modalities. Moreover, it provides a relative quality index for each method/dataset to give insights into the generation process and reported performance. This thorough literature review and evaluation of existing computer graphics techniques for synthetic training data is a comprehensive reference that is useful to inspect the various trends and alternatives and detect missing attributes and future directions.
2. Paper **B** presents a novel computer graphics-based data generation system capable of producing large-scale, highly realistic, content-rich, and diverse image sets for visual scene understanding of natural images. This work contributes with a pipeline constructed of a procedural world modeling engine and state-of-the-art light transport simulation using path-tracing techniques. This approach yields an exponential explosion of potential worlds where each image is essentially a unique virtual setup and highly different from the rest of the generated images. It is a flexible and scalable system that provides increased levels of control over the generation process and comes with per-pixel accurate annotations for semantic segmentation. The evaluation analysis showed that realism, defined as accurate scene definition, geometry, light transport simulation, materials, textures, and

optical effects, has a positive impact on performance, producing state-of-the-art, at the time, results.

3. Paper **F** focuses on GAN-based synthetic training data generation that can be used in isolation from real data, and operate for example, as an anonymization method of histopathology images. This paper contributes empirical evidence for the importance of using GANs in ensembles, especially in an independently trained scheme. This is a simple but powerful approach to improve the quality and diversity of the generated image dataset, especially for complex data and simple GAN variants. Compared to more sophisticated boosting techniques, this scheme lowers the risk of overfitting to specific image modes, caused by focusing on a narrow part of the data distribution, a situation that is not acceptable for anonymization purposes.

6.1.2 Data augmentation

For data augmentation, the thesis contributes to practices that utilize generative modeling. The most promising augmentation technique proved to be inter-domain transformations through CycleGANs. This observation stands for both histopathology and natural image modalities, tested for binary and multi-label classification tasks. In this training mode, the presented contributions focus on scarce training sets, which in some cases may also exhibit imbalanced class representations. These sets pose as extreme training data insufficient cases, where the class sizes range from a few hundred to a few thousand images, and proposes solutions to improve both model performance and stability/robustness. Specifically,

1. Papers **C** and **D** contribute to a comprehensive augmentations scheme for low-data colon cancer lymphatic metastasis detection. Paper **C** was the first to introduce the idea of primary colon cancer tumor (intra-organ) and other organs' tissue (inter-organ) augmentations, while paper **D** is the first large-scale study on intra- and inter-organ data augmentations in digital pathology for metastasis detection. Three availability scenarios of the target domain are studied, one of which completely excludes tissue from metastatic lymph node colon cancer, and two publicly available annotated sets of different tissue and cancer types are employed. The papers provide evidence that it is possible to improve both model performance and robustness using the proposed augmentations either in a direct mix fashion or by generating synthetic samples through CycleGAN transformations, especially when the representation shift between the added and target domain is large.
2. Finally, paper **E** expands the intra- and inter-domain GAN-based data generation for augmentation purposes to natural images, showcased for single-image multi-label weather classification. The main contributions are a thorough investigation and evaluation of GAN-generated augmentations compared to classical image processing transformations and practical proof that CycleGAN inter-domain synthetic data can be a viable augmentation solution to very scarce and imbalanced

natural image sets, too. These results contribute to a less developed/explored research area because this modality is quite different from medical imaging samples and can be challenging for GANs due to its high diversity.

6.2 Discussion

The first part of this book has described, analyzed, and discussed different aspects and formulations of image synthesis methods and synthetic visual data suitable for supervised deep learning. By now, the reader may have wondered what the differences are between image synthesis and image manipulation/processing. How can we delimit the line between them and define hard rules between two terms that can widely overlap? And although computer graphics-generated imagery produces visual outcomes that are more well-defined and intuitive, the question becomes more prominent for GAN-based image generation. One way to approach this, and define what can be classified as a new, synthetic image, is to look at its generation process. For example, low-level image manipulations, that can be easily *inverted* and do not contribute to the image space with new information, like geometric transformations, cannot be considered as synthetic data generation mechanisms. In contrast, generative image modeling is a learning process that finds functions to draw samples from a data distribution or map one domain to another. This process may not add new content to the generated images, in the sense of new 3D scene attributes, but may add new information to the image space. Therefore, new synthetic image data emerge. This dissertation and its enclosed papers employ this conceptual separation between image processing and synthesis.

On a general and concluding note, synthetic data can significantly benefit deep learning-driven computer vision in several ways. The first and foremost advantage concerns the automatic ground truth generation. Especially for annotation-expensive image types and tasks, synthetic data provide a viable alternative while, depending on the generation pipeline, it is possible to output several kinds of ground truth types. This way various problems can utilize synthetic training data. Moreover, image synthesis comes with increased control over the generation process. Compared to capturing procedures, data generation enables large-scale image production where the content, diversity and complexity can vary at a lower effort cost.

Finally, the more this work supports data generation development and the use of synthetic data, the more it recognizes the need for well-defined synthesis and utilization frameworks. Synthetic image sets for machine learning can ensure both data diversity and privacy protection, especially when personal information is involved. They are also capable of integrating the worlds variations, potentially directly in the generation pipeline, and therefore they are capable of reducing the bias introduced in the involved sets. Whenever applicable and under the general societal and ethical norms, this thesis stands that the use of synthetic data, in a legal and human-rights preserving manner, can contribute to the democratization of computer vision. Synthetic data calls for strategies

that depict reality and provide enough variations of the majority and minority relevant subjects and scenarios.

In the last two decades, the idea of synthetic data for visual machine learning has been steadily rising. This is reflected in Figure 7.1 by the increasing fraction of publications connected to synthetic data, i.e. when the number of publications is normalized by the total number of machine learning papers. The shown fraction concerns published machine learning papers per year that include in their title or abstract the words “*synthetic*” and “*data*”. Naturally, there can be papers that fit the search criteria but do not deal with synthetic data for visual machine learning, and vice versa papers that treat synthetic training data but do not use the specified terms in their title/abstract. Thus, the depicted numbers are not absolute but able to reveal the general trend and indicate further research growth.

Even though some computer vision tasks share similar qualities and can be served by the same datasets, not all applications have similar needs. Recognizing this, what is probably missing at the moment are unified guidelines for synthetic data generation, targeted to tasks families that seem to respond well to specific characteristics. For example, there are certain fields, like optical flow and depth and disparity estimation, where image diversity is desirable over realism. The same does not hold for other groups of tasks, like recognition, where both diversity and realism are a must.

Researching the relevant fields during this thesis, it became evident that optimal synthesis configurations with perfect combinations do not exist yet. For example, a possibly highly efficient data generation pipeline would scale easily and create feature varying, diverse, and, in cases, highly realistic images fast and automatically, with no manual effort in the various production and annotation stages. In terms of computer graphics, this translates to employing more extensively procedural modeling strategies, accurate light transport simulation — possibly carried by modern GPUs or distributed to cloud services, and some levels of domain and rendering randomization. On the other hand, GAN-based image synthesis could leverage the wide range of the existing and already available visual data to construct new representations for various purposes. Inter-domain image synthesis is a promising tactic capable of facilitating more tasks than the ones the source domain was designed for. Additionally, the generation process could potentially be accelerated and stabilized by incorporating progressive growing or super-resolution schemes into the learning process. However, the most intuitive and prominent future direction is the development and more extensive use of hybrid data

Figure 7.1: The fraction of published machine learning papers per year with “*synthetic*” and “*data*” in the paper title/abstract, according to the ScienceDirect database.

generation methods. Exploiting elements from both computer graphics and generative modeling seems like the next step to image synthesis for visual machine learning.

Given the current, ongoing, and continuous growth of deep learning and relevant data-centric practices, it is no exaggeration to assume that synthetic data is almost certain to be integral part of the learning process. The better their quality becomes, the more they will be used, and they will play even more of a central role in visual machine learning. This applies to both training and testing sets. Future development points to increased research around image synthesis for optimizing the training data distribution. For example, when real data capture the main mode of the distribution and synthetic data model the rare (and even unrealistic) samples. Synthetic visual data can also play an important role when evaluating models. Their ability on feature variation and meta-data generation opens up possibilities in detailed statistical testing across different kinds of parameters.

Their use can be expanded to less obvious applications than the well-known and tested areas by the computer vision research community. To name only a few, agriculture, quality inspection of automated manufacturing processes, and essential consumable products like food and drinking water or predictive maintenance pipelines can greatly benefit from synthetic data and deep learning solutions. In many of these areas, the image complexity is limited and, thus, the generation process can be framed under structured parameters.

With the gaze to the future, this thesis completes the circle of a study of image synthesis methods and applications for deep learning. Computer-generated images will be, if not already are at some stage and extent, an integral part of the learning process. Computer graphics and generative models can work synergistically and apart in this direction. Well-defined synthetic data that respect scientific values and individuals’ rights have

indeed the potential to advance research and real-life applicability.

Bibliography

- [1] FDA Authorizes Software that Can Help Identify Prostate Cancer, 2021. [Online]. Available: <https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer>. [page 15]
- [2] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey. Using Synthetic Data to Improve Facial Expression Analysis With 3D Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. [page 24]
- [3] A. Abdelhamed, M. A. Brubaker, and M. S. Brown. Noise Flow: Noise Modeling With Conditional Normalizing Flows. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. [page 24]
- [4] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. [page 40]
- [5] T. Akenine-Möller, E. Haines, and N. Hoffman. *Real-Time Rendering*. AK Peters/crc Press, 2019. [page 27]
- [6] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. [page 70]
- [7] S. Al-Janabi, A. Huisman, and P. J. Van Diest. Digital pathology: current status and future perspectives. *Histopathology*, 61(1):1–9, 2012. [page 25]
- [8] G. Algan and I. Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021. [page 52]
- [9] A. Ali-Gombe, E. Elyan, and C. Jayne. Multiple fake classes gan for data augmentation in face image dataset. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. [page 71]
- [10] J. L. Alireza Shafaei and M. Schmidt. Play and Learn: Using Video Games to Train Computer Vision Models. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 26.1–26.13. BMVA Press, 2016. [pages 24 and 58]

- [11] G. Amato, L. Ciampi, F. Falchi, C. Gennaro, and N. Messina. Learning Pedestrian Detection from Virtual Worlds. In E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 302–312, Cham, 2019. Springer International Publishing. [page 24]
- [12] M. Angus, M. ElBalkini, S. Khan, A. Harakeh, O. Andrienko, C. Reading, S. Waslander, and K. Czarnecki. Unlimited road-scene synthetic annotation (ursa) dataset. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 985–992. IEEE, 2018. [page 24]
- [13] A. Antoniou, A. Storkey, and H. Edwards. Data Augmentation Generative Adversarial Networks. *arXiv preprint arXiv:1711.04340*, 2017. [page 71]
- [14] A. Appel. Some Techniques for Shading Machine Renderings of Solids. In *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference, AFIPS '68 (Spring)*, pages 37–45, New York, NY, USA, 1968. Association for Computing Machinery. [page 37]
- [15] R. B. Arantes, G. Vogiatzis, and D. R. Faria. CSC-GAN: Cycle and Semantic Consistency for Dataset Augmentation. In G. Bebis, Z. Yin, E. Kim, J. Bender, K. Subr, B. C. Kwon, J. Zhao, D. Kalkofen, and G. Baciu, editors, *Advances in Visual Computing*, pages 170–181, Cham, 2020. Springer International Publishing. [page 71]
- [16] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017. [pages 44 and 45]
- [17] M. Ashikhmin and P. Shirley. An Anisotropic Phong BRDF Model. *Journal of Graphics Tools*, 5(2):25–32, 2000. [page 29]
- [18] H.-J. Bae, C.-W. Kim, N. Kim, B. Park, N. Kim, J. B. Seo, and S. M. Lee. A Perlin Noise-Based Augmentation Strategy for Deep Learning with Small Data Samples of HRCT Images. *Scientific Reports*, 8(1), 2018. [page 69]
- [19] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011. [page 24]
- [20] A. S. Baslamisli, T. T. Groenestegge, P. Das, H.-A. Le, S. Karaoglu, and T. Gevers. Joint learning of intrinsic images and semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, 2018. [page 24]
- [21] A. S. Baslamisli, H.-A. Le, and T. Gevers. CNN Based Learning Using Reflection and Retinex Models for Intrinsic Image Decomposition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [page 24]

- [22] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. [page 59]
- [23] S. Beigpour, M. Serra, J. van de Weijer, R. Benavente, M. Vanrell, O. Penacchio, and D. Samaras. Intrinsic image evaluation on synthetic complex scenes. In *2013 IEEE International Conference on Image Processing*, pages 285–289. IEEE, 2013. [page 24]
- [24] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy Layer-Wise Training of Deep Networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007. [page 6]
- [25] Y. Bengio, E. r. Thibodeau-Laufer, G. Alain, and J. Yosinski. Deep Generative Stochastic Networks Trainable by Backprop. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32 of ICML'14, pages II226–II234. JMLR.org, 2014. [page 41]
- [26] U. Bergmann, N. Jetchev, and R. Vollgraf. Learning Texture Manifolds with the Periodic Spatial GAN. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 469–477. PMLR, 2017. [page 59]
- [27] S. Bi, K. Sunkavalli, F. Perazzi, E. Shechtman, V. G. Kim, and R. Ramamoorthi. Deep CG2Real: Synthetic-to-Real Translation via Image Disentanglement. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. [page 24]
- [28] O. Bielova, R. Hansch, A. Ley, and O. Hellwich. A Digital Image Processing Pipeline for Modelling of Realistic Noise in Synthetic Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. [page 24]
- [29] A. Birhane and F. Cummins. Algorithmic Injustices: Towards a Relational Ethics. In *NeurIPS Workshop on Black in AI*, 2019. [page 55]
- [30] A. Birhane and V. U. Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, 2021. [pages 51 and 55]
- [31] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. [page 40]
- [32] M. Bikowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. [page 44]
- [33] H. Blasinski, J. Farrell, T. Lian, Z. Liu, and B. Wandell. Optimizing Image Acquisition Systems for Autonomous Driving. *Electronic Imaging*, 2018(5):161–1–161–7, 2018. [page 24]

- [34] J. F. Blinn. Models of Light Reflection for Computer Synthesized Pictures. In *Proceedings of the 4th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '77, page 192198, New York, NY, USA, 1977. Association for Computing Machinery. [page 29]
- [35] A. Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. [page 45]
- [36] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert. Gan Augmentation: Augmenting Training Data using Generative Adversarial Networks. *arXiv preprint arXiv:1810.10863*, 2018. [pages 71 and 75]
- [37] J. Brierley, M. Gospodarowicz, and C. Wittekind, editors. *UICC TNM Classification of Malignant Tumours*. Wiley-Blackwell, Chichester, 8th edition, 2016. [page 73]
- [38] A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019. [page 59]
- [39] A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019. [pages 41, 47, and 59]
- [40] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recogn. Lett.*, 30(2):88–97, 2009. [pages 2 and 58]
- [41] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. [pages 24 and 58]
- [42] P. Bándi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. Çetin, E. Halc, H. Jackson, R. Chen, F. Both, J. Franke, H. Küsters-Vandeveld, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. [page 25]
- [43] D. A. Calian, F. Stimberg, O. Wiles, S.-A. Rebuffi, A. Gyorgy, T. Mann, and S. Gowal. Defending Against Image Corruptions Through Adversarial Augmentations. *arXiv preprint arXiv:2104.01086*, 2021. [page 70]
- [44] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing Training Images for Boosting Human 3D Pose Estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. [pages 24 and 58]

- [45] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29 of *NIPS'16*, page 21802188, Red Hook, NY, USA, 2016. Curran Associates Inc. [page 59]
- [46] K. Chmielinski, S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, and Y. C. Qiu. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. In *NeurIPS Workshop on Dataset Curation and Security*, 2020. [page 52]
- [47] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. [page 47]
- [48] F. Chollet. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [page 54]
- [49] P. A. Cicalese, A. Mobiny, P. Yuan, J. Becker, C. Mohan, and H. V. Nguyen. StyPath: Style-Transfer Data Augmentation for Robust Histology Image Classification. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 351–361, Cham, 2020. Springer International Publishing. [page 71]
- [50] R. L. Cook and K. E. Torrance. A Reflectance Model for Computer Graphics. *SIGGRAPH Comput. Graph.*, 15(3):307316, 1981. [page 29]
- [51] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. [pages 2, 58, 60, and 62]
- [52] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. End-to-End Adversarial Retinal Image Synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017. [pages 24 and 59]
- [53] K. Crawford and T. Paglen. Excavating AI: The Politics of Images in Machine Learning Training Sets, 2019. [Online]. Available: <https://excavating.ai/>. [pages 51 and 53]
- [54] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. AutoAugment: Learning Augmentation Strategies From Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [page 69]

- [55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. [page 54]
- [56] T. DeVries and G. W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017. [page 69]
- [57] N. Dimitriou, O. Arandjelovi, and P. D. Caie. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine*, 6:264, 2019. [page 25]
- [58] L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear Independent Components Estimation. *arXiv preprint arXiv:1410.8516*, 2014. [pages 6 and 40]
- [59] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017. [pages 6 and 40]
- [60] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow With Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. [pages 24 and 58]
- [61] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In S. Levine, V. Vanhoucke, and K. Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2017. [page 58]
- [62] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 2017. [page 25]
- [63] G. Eilertsen, S. Hajisharif, P. Hanji, A. Tsirikoglou, R. K. Mantiuk, and J. Unger. How to cheat with metrics in single-image HDR reconstruction. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2021.
- [64] G. Eilertsen, A. Tsirikoglou, C. Lundström, and J. Unger. Ensembles of GANs for Synthetic Training Data Generation. In *ICLR Workshop on Synthetic Data Generation (SDG)*. arXiv, 2021. [page xi]
- [65] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara. Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World. In *The European Conference on Computer Vision (ECCV)*, 2018. [pages 24 and 58]
- [66] S. E. Fahlman, G. E. Hinton, and T. J. Sejnowski. Massively Parallel Architectures for AI: Netl, Thistle, and Boltzmann Machines. In *Proceedings of the Third AAAI Conference on Artificial Intelligence, AAAI’83*, pages 109–113. AAAI Press, 1983. [page 40]

- [67] N. Farahani, A. V. Parwani, and L. Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33, 2015. [page 25]
- [68] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard. Adaptive data augmentation for image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3688–3692, 2016. [page 67]
- [69] J. D. Foley, F. D. Van, A. Van Dam, S. K. Feiner, J. F. Hughes, and J. Hughes. *Computer graphics: principles and practice*, volume 12110. Addison-Wesley Professional, 1996. [pages 4 and 27]
- [70] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. [page 71]
- [71] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. VirtualWorlds as Proxy for Multi-Object Tracking Analysis. In *CVPR*, 2016. [pages 24, 57, and 58]
- [72] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. [page 70]
- [73] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. Datasheets for Datasets. *Commun. ACM*, 64(12):8692, 2021. [pages 52 and 53]
- [74] B. Gecer, B. Bhattarai, J. Kittler, and T.-K. Kim. Semi-Supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model. In *The European Conference on Computer Vision (ECCV)*, 2018. [pages 24 and 59]
- [75] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang. Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, page 325336, New York, NY, USA, 2020. Association for Computing Machinery. [page 53]
- [76] M. Germain, K. Gregor, I. Murray, and H. Larochelle. MADE: Masked Autoencoder for Distribution Estimation. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 881–889, Lille, France, 2015. PMLR. [pages 6 and 40]
- [77] C. Gong, T. Ren, M. Ye, and Q. Liu. MaxUp: Lightweight Adversarial Training With Data Augmentation Improves Neural Network Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2474–2483, 2021. [page 70]

- [78] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. [pages 4 and 41]
- [79] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>. [pages 7 and 40]
- [80] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, Conference Track Proceedings*, 2015. [page 70]
- [81] C. M. Goral, K. E. Torrance, D. P. Greenberg, and B. Battaile. Modeling the Interaction of Light between Diffuse Surfaces. *SIGGRAPH Comput. Graph.*, 18(3): 213–222, 1984. [page 35]
- [82] H. Gouraud. Continuous Shading of Curved Surfaces. *IEEE Transactions on Computers*, C-20(6):623–629, 1971. [page 29]
- [83] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(null):723773, 2012. [page 44]
- [84] A. Grover and S. Ermon. Boosted Generative Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. [page 65]
- [85] A. Gupta, S. Venkatesh, S. Chopra, and C. Ledig. Generative Image Translation for Data Augmentation of Bone Lesion Pathology. In M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, and T. Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 225–235. PMLR, 2019. [page 71]
- [86] S. Hajisharif. *Computational Photography: High Dynamic Range and Light Fields*. PhD dissertation, Linköping University Electronic Press, 2020. [page 2]
- [87] V. Haltakov, C. Unger, and S. Ilic. Framework for Generation of Synthetic Ground Truth Data for Driver Assistance Applications. In J. Weickert, M. Hein, and B. Schiele, editors, *Pattern Recognition*, pages 323–332. Springer Berlin Heidelberg, 2013. [page 24]
- [88] C. Han, K. Murao, T. Noguchi, Y. Kawata, F. Uchiyama, L. Rundo, H. Nakayama, and S. Satoh. Learning More with Less: Conditional PGGAN-Based Data Augmentation for Brain Metastases Detection Using Highly-Rough Annotation on MR Images. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, page 119127, New York, NY, USA, 2019. Association for Computing Machinery. [page 71]

- [89] A. Handa, V. Pătrăucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Understanding Real World Indoor Scenes With Synthetic Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [pages 24 and 58]
- [90] A. Handa, V. Pătrăucean, S. Stent, and R. Cipolla. SceneNet: An Annotated Model Generator for Indoor Scene Understanding. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5737–5743. IEEE, 2016. [pages 24 and 58]
- [91] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. [pages 16 and 17]
- [92] J. Hedlund, A. Eklund, and C. Lundström. Key insights in the AIDA community policy on sharing of clinical imaging data for research in Sweden. *Scientific Data*, 7(1), 2020. [page 53]
- [93] D. Hendrycks*, N. Mu*, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. AugMix: A Simple Method to Improve Robustness and Uncertainty under Data Shift. In *International Conference on Learning Representations*, 2020. [page 70]
- [94] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. [page 70]
- [95] D. Hernandez-Juarez, L. Schneider, A. Espinosa, D. Vazquez, A. M. Lopez, U. Franke, M. Pollefeys, and J. C. Moure. Slanted Stixels: Representing San Francisco’s Steepest Streets. In *British Machine Vision Conference (BMVC)*, 2017, 2017. [pages 24 and 57]
- [96] 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. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. [page 44]
- [97] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. [page 54]
- [98] G. E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11(10):428–434, 2007. [page 6]
- [99] G. E. Hinton, S. Osindero, and Y.-W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006. [page 6]
- [100] T. Hodaň, V. Vineet, R. Gal, E. Shalev, J. Hanzelka, T. Connell, P. Urbina, S. N. Sinha, and B. Guenter. Photorealistic Image Synthesis for Object Instance Detection. In

- 2019 *IEEE International Conference on Image Processing (ICIP)*, pages 66–70. IEEE, 2019. [page 24]
- [101] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018. [page 52]
- [102] M. Hong, J. Choi, and G. Kim. StyleMix: Separating Content and Style for Enhanced Data Augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14862–14870, 2021. [pages 70 and 71]
- [103] Y. Hong, U. Hwang, J. Yoo, and S. Yoon. How Generative Adversarial Networks and Their Variants Work: An Overview. *ACM Comput. Surv.*, 52(1), 2019. [page 39]
- [104] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz. Robust Histopathology Image Analysis: To Label or to Synthesize? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8525–8534, 2019. [pages 24 and 66]
- [105] R. Huang, S. Zhang, T. Li, and R. He. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2458–2467, 2017. [page 59]
- [106] X. Huang and S. Belongie. Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [page 70]
- [107] B. Hurl, K. Czarnecki, and S. Waslander. Precise Synthetic Image and LiDAR (PreSIL) Dataset for Autonomous Vehicle Perception. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2522–2529, 2019. [page 24]
- [108] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. FAccT ’21, page 560575, New York, NY, USA, 2021. Association for Computing Machinery. [page 53]
- [109] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [page 24]
- [110] H. Inoue. Data Augmentation by Pairing Samples for Images Classification. *arXiv preprint arXiv:1801.02929*, 2018. [page 69]
- [111] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. [pages 46, 59, and 60]

- [112] P. T. Jackson, A. Atapour-Abarghouei, S. Bonner, T. P. Breckon, and B. Obara. Style Augmentation: Data Augmentation via Style Randomization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. [pages 70 and 71]
- [113] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(29), 2016. [page 25]
- [114] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3):101493, 2020. [page 54]
- [115] H. W. Jensen. Global Illumination using Photon Maps. In X. Pueyo and P. Schröder, editors, *Rendering Techniques '96*, pages 21–30, Vienna, 1996. Springer Vienna. [page 37]
- [116] H. W. Jensen. *Realistic image synthesis using photon mapping*. A K Peters/CRC Press, 2001. [page 37]
- [117] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. Neural Style Transfer: A Review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3365–3385, 2020. [page 70]
- [118] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. [page 70]
- [119] M. Johnson-Roberson, C. Barto, R. C. S. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan. Driving in the Matrix: Can virtual worlds replace human-generated annotations for real world tasks? *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753, 2016. [pages 24 and 58]
- [120] J. T. Kajiya. The rendering equation. *SIGGRAPH Comput. Graph.*, 20(4):143–150, 1986. [pages 32 and 37]
- [121] G. Kang, X. Dong, L. Zheng, and Y. Yang. Patchshuffle Regularization. *arXiv preprint arXiv:1707.07103*, 2017. [page 69]
- [122] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler. Meta-Sim: Learning to Generate Synthetic Datasets. In *ICCV*, 2019. [pages 24 and 60]
- [123] D. Karimi, G. Nir, L. Fazli, P. C. Black, L. Goldenberg, and S. E. Salcudean. Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1413–1426, 2020. [page 68]

- [124] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*, 2018. [pages 41, 45, 47, and 59]
- [125] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. [pages 45 and 47]
- [126] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020. [pages 41, 59, and 75]
- [127] A. Khan, A. Sohail, U. Zahoor, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020. [page 17]
- [128] S. Khan, B. Phan, R. Salay, and K. Czarnecki. ProcSy: Procedural Synthetic Dataset Generation Towards Influence Factor Studies Of Semantic Segmentation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–96, 2019. [page 24]
- [129] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865. PMLR, 2017. [pages 47 and 59]
- [130] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. [page 18]
- [131] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations (ICLR)*, 2014. [pages 6 and 40]
- [132] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved Variational Inference with Inverse Autoregressive Flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. [page 40]
- [133] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter. Training Deep Face Recognition Systems with Synthetic Data. *arXiv preprint arXiv:1802.05891*, 2018. [page 24]
- [134] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. [page 54]
- [135] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q.

- Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. [pages 16, 17, and 68]
- [136] J. Kronander. *Physically Based Rendering of Synthetic Objects in Real Environments*. PhD thesis, Linköping UniversityLinköping University, Media and Information Technology, Faculty of Science & Engineering, 2015. [page 35]
- [137] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, T. Duerig, and V. Ferrari. The Open Images Dataset V4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. [page 54]
- [138] H. Larochelle and I. Murray. The Neural Autoregressive Distribution Estimator. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 29–37, Fort Lauderdale, FL, USA, 2011. PMLR. [pages 6 and 40]
- [139] Y. Lecun. *Generalization and network design strategies*. Elsevier, 1989. [page 16]
- [140] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989. [page 16]
- [141] Y. LeCun, S. Chopra, R. Hadsell, F. J. Huang, and et al. A Tutorial on Energy-Based Learning. In *PREDICTING STRUCTURED DATA*. MIT Press, 2006. [page 6]
- [142] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. [page 7]
- [143] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. [page 59]
- [144] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. [page 44]
- [145] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2Noise: Learning Image Restoration without Clean Data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2965–2974, Stockholmsmässan, Stockholm Sweden, 2018. PMLR. [page 24]
- [146] C. Li and M. Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 702–716, Cham, 2016. Springer International Publishing. [page 59]

- [147] Y. Li, K. Swersky, and R. Zemel. Generative Moment Matching Networks. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727, Lille, France, 2015. PMLR. [page 41]
- [148] Z. Li and N. Snavely. CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering. In *The European Conference on Computer Vision (ECCV)*, 2018. [page 24]
- [149] K. Lindman, J. F. Rose, M. Lindvall, and C. B. Stadler. Skin data from the Visual Sweden project DROID, 2019, <http://doi:10.23698/aida/drsk>. [page 25]
- [150] M.-Y. Liu and O. Tuzel. Coupled Generative Adversarial Networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. [page 45]
- [151] S. Liu, J. Zhang, Y. Chen, Y. Liu, Z. Qin, and T. Wan. Pixel Level Data Augmentation for Semantic Image Segmentation Using Generative Adversarial Networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1902–1906, 2019. [page 71]
- [152] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng, and M. C. Stumpe. Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv preprint arXiv:1703.02442*, 2017. [page 2]
- [153] Z. Liu, T. Lian, J. Farrell, and B. Wandell. Soft Prototyping Camera Designs for Car Detection Based on a Convolutional Neural Network. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019. [page 24]
- [154] Z. Liu, M. Shen, J. Zhang, S. Liu, H. Blasinski, T. Lian, and B. Wandell. A system for generating complex physically accurate sensor images for automotive applications. *Electronic Imaging*, 2019(15):53–1–53–6, 2019. [page 24]
- [155] D. Lopez-Paz and M. Oquab. Revisiting Classifier Two-Sample Tests. In *ICLR (Poster)*. OpenReview.net, 2017. [page 44]
- [156] J. Löw, J. Kronander, A. Ynnerman, and J. Unger. BRDF Models for Accurate and Efficient Rendering of Glossy Surfaces. *ACM Trans. Graph.*, 31(1), 2012. [page 29]
- [157] S. Lutz, K. Amliantis, and A. Smolic. AlphaGAN: Generative adversarial networks for natural image matting. In *British Machine Vision Conference (BMVC)*, page 259. BMVA Press, 2018. [page 59]
- [158] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi. BAGAN: Data Augmentation with Balancing GAN. *arXiv preprint arXiv:1803.09655*, 2018. [page 71]

- [159] J. Marin, D. Vázquez, D. Gerónimo, and A. M. López. Learning appearance in virtual scenarios for pedestrian detection. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 137–144. IEEE, 2010. [page 24]
- [160] S. Marschner and P. Shirley. *Fundamentals of Computer Graphics*. CRC Press, 2018. [page 27]
- [161] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [pages 24 and 58]
- [162] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox. What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation? *International Journal of Computer Vision*, 126:942–960, 2018. [page 2]
- [163] A. Mayor. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton University Press, 2020. [page 6]
- [164] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-Training on Indoor Segmentation? In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. [page 24]
- [165] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. [page 6]
- [166] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. [page 24]
- [167] M. Miceli, T. Yang, L. Naudts, M. Schuessler, D. Serbanescu, and A. Hanna. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 161172, New York, NY, USA, 2021. Association for Computing Machinery. [page 52]
- [168] A. Mikoajczyk and M. Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122, 2018. [pages 67 and 71]
- [169] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014. [page 45]
- [170] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. [page 70]

- [171] S. Mohamed and B. Lakshminarayanan. Learning in Implicit Generative Models. *arXiv preprint arXiv:1610.03483*, 2016. [page 6]
- [172] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 86–94, 2017. [page 70]
- [173] M. Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem. Sim4CV: A Photo-Realistic Simulator for Computer Vision Applications. *International Journal of Computer Vision*, 126(9):902–919, 2018. [page 58]
- [174] P. Müller, P. Wonka, S. Haegler, A. Ulmer, and L. Van Gool. Procedural Modeling of Buildings. In *ACM SIGGRAPH 2006 Papers, SIGGRAPH '06*, page 614623, New York, NY, USA, 2006. Association for Computing Machinery. [page 31]
- [175] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kotschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [page 60]
- [176] F. E. Nicodemus. Directional reflectance and emissivity of an opaque surface. *Appl. Opt.*, 4(7):767–775, 1965. [page 33]
- [177] S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. [pages 44 and 45]
- [178] A. Odena, C. Olah, and J. Shlens. Conditional Image Synthesis with Auxiliary Classifier GANs. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR, 2017. [page 44]
- [179] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, pages 4797–4805, Red Hook, NY, USA, 2016. Curran Associates Inc. [page 40]
- [180] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng. Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access*, 7:36322–36333, 2019. [pages 39 and 45]
- [181] G. Papamakarios, T. Pavlakou, and I. Murray. Masked Autoregressive Flow for Density Estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pages 2335–2344, Red Hook, NY, USA, 2017. Curran Associates Inc. [page 40]

- [182] D. Park and D. Ramanan. Articulated pose estimation with tiny synthetic videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2015. [pages 24 and 58]
- [183] E. Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union L*, 119(1), 2016. [page 53]
- [184] L. Perez and J. Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv preprint arXiv:1712.04621*, 2017. [pages 67 and 71]
- [185] M. Pharr, W. Jakob, and G. Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2016. [pages 4, 27, and 35]
- [186] B. T. Phong. Illumination for Computer Generated Pictures. *Commun. ACM*, 18(6):311317, 1975. [page 29]
- [187] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012. [pages 24 and 58]
- [188] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3309–3318, 2017. [pages 62 and 63]
- [189] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield. Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data. *arXiv preprint arXiv:1810.10093*, 2018. [page 24]
- [190] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, and Y. Wang. UnrealCV: Virtual Worlds for Computer Vision. In *Proceedings of the 25th ACM International Conference on Multimedia, MM '17*, pages 1221–1224. ACM, 2017. [page 24]
- [191] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations (ICLR)*, 2016. [page 43]
- [192] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations (ICLR)*, 2016. [page 45]

- [193] R. Ramanath, W. Snyder, Y. Yoo, and M. Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43, 2005. [page 2]
- [194] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014. [page 68]
- [195] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 2016. PMLR. [page 59]
- [196] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars. Deep Reflectance Maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [page 24]
- [197] T. Remez, J. Huang, and M. Brown. Learning to Segment via Cut-and-Paste. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. [pages 24 and 60]
- [198] D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 2015. PMLR. [page 40]
- [199] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Beijing, China, 2014. PMLR. [page 40]
- [200] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for Data: Ground Truth from Computer Games. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 102–118, Cham, 2016. Springer International Publishing. [pages 24, 57, 58, and 63]
- [201] S. R. Richter, Z. Hayder, and V. Koltun. Playing for Benchmarks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. [pages 24 and 57]
- [202] S. R. Richter, H. A. AlHaija, and V. Koltun. Enhancing Photorealism Enhancement. *arXiv:2105.04619*, 2021. [pages 24 and 60]
- [203] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. [pages 24, 38, 57, 58, and 63]

- [204] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. [page 6]
- [205] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. [pages 6, 17, and 18]
- [206] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986. [page 17]
- [207] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. [page 54]
- [208] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671*, 2016. [page 47]
- [209] R. Salakhutdinov and G. Hinton. Deep Boltzmann Machines. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009. PMLR. [page 40]
- [210] B. E. A. Saleh and M. C. Teich. *Fundamentals of Photonics*. John Wiley & sons, 2019. [page 32]
- [211] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved Techniques for Training GANs. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. [page 44]
- [212] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. PixelCNN++: A Pixel-CNN Implementation with Discretized Logistic Mixture Likelihood and Other Modifications. In *International Conference on Learning Representations (ICLR)*, 2017. [page 40]
- [213] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, 9(1), 2019. [page 71]
- [214] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. [page 58]
- [215] D. Saxena and J. Cao. Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Comput. Surv.*, 54(3), 2021. [page 39]

- [216] M. K. Scheuerman, A. Hanna, and E. Denton. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), 2021. [pages 51 and 53]
- [217] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young. Machine Learning: The High Interest Credit Card of Technical Debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014. [page 53]
- [218] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden Technical Debt in Machine Learning Systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. [page 53]
- [219] S. Shah, D. Dey, C. Lovett, and A. Kapoor. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In M. Hutter and R. Siegwart, editors, *Field and Service Robotics*, pages 621–635, Cham, 2018. Springer International Publishing. [page 58]
- [220] J. Shijie, W. Ping, J. Peiyi, and H. Siping. Research on data augmentation for image classification based on convolution neural networks. In *2017 Chinese Automation Congress (CAC)*, pages 4165–4170, 2017. [pages 67 and 69]
- [221] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In A. Gooya, O. Goksel, I. Oguz, and N. Burgos, editors, *Simulation and Synthesis in Medical Imaging*, pages 1–11, Cham, 2018. Springer International Publishing. [pages 24, 59, and 71]
- [222] C. Shorten and T. M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 2019. [pages 67 and 69]
- [223] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. IEEE, 2011. [pages 24 and 58]
- [224] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from Simulated and Unsupervised Images through Adversarial Training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2242–2251, 2017. [pages 24 and 60]
- [225] H. A. Sial, R. Baldrich, and M. Vanrell. Deep intrinsic decomposition trained on surreal scenes yet with realistic light effects. *J. Opt. Soc. Am. A*, 37(1):1–15, 2020. [pages 24 and 58]

- [226] P. Simard, D. Steinkraus, and J. Platt. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963, 2003. [page 68]
- [227] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. [pages 16 and 17]
- [228] Y. Song and D. P. Kingma. How to Train Your Energy-Based Models. *arXiv preprint arXiv:2101.03288*, 2021. [page 6]
- [229] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008. [page 54]
- [230] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström. A Closer Look at Domain Shift for Deep Learning in Histopathology. *arXiv preprint arXiv:1909.11575*, 2019. [page 68]
- [231] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, 2021. [page 68]
- [232] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for CNN: Viewpoint Estimation in Images Using CNNs Trained With Rendered 3D Model Views. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. [pages 24 and 58]
- [233] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. [page 70]
- [234] C. Summers and M. J. Dinneen. Improved Mixed-Example Data Augmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1262–1270, 2019. [page 69]
- [235] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. [pages 16 and 17]
- [236] R. Takahashi, T. Matsubara, and K. Uehara. Data Augmentation Using Random Image Cropping and Patching for Deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2020. [page 69]
- [237] L. Taylor and G. Nitschke. Improving Deep Learning with Generic Data Augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547, 2018. [pages 67, 68, and 69]

- [238] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30. IEEE, 2017. [pages 24 and 58]
- [239] Y. Tokozume, Y. Ushiku, and T. Harada. Between-Class Learning for Image Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5494, 2018. [page 69]
- [240] I. O. Tolstikhin, S. Gelly, O. Bousquet, C.-J. SIMON-GABRIEL, and B. Schölkopf. AdaGAN: Boosting Generative Models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. [page 65]
- [241] D. Tome, P. Peluse, L. Agapito, and H. Badino. xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. [page 24]
- [242] T. Tongbuasirilai, J. Unger, J. Kronander, and M. Kurt. Compact and intuitive data-driven BRDF models. *The Visual Computer*, 36(4):855–872, 2020. [page 29]
- [243] A. Torralba, R. Fergus, and W. T. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008. [page 54]
- [244] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield. Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 969–977, 2018. [pages 24 and 57]
- [245] J. Tremblay, T. To, and S. Birchfield. Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. [pages 24, 57, and 58]
- [246] A. Triastcyn and B. Faltings. Generating Artificial Data for Private Deep Learning. *arXiv preprint arXiv:1803.03148*, 2018. [pages 24 and 59]
- [247] D. S. Trigueros, L. Meng, and M. Hartnett. Generating Photo-Realistic Training Data to Improve Face Recognition Accuracy. *arXiv preprint arXiv:1811.00112*, 2018. [pages 24 and 59]
- [248] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*, 2019. [page 70]

- [249] A. Tsirikoglou, J. Kronander, P. Larsson, T. Tongbuasirilai, A. Gardner, and J. Unger. Differential Appearance Editing for Measured BRDFs. In *ACM SIGGRAPH 2016 Talks*, SIGGRAPH '16, New York, NY, USA, 2016. Association for Computing Machinery.
- [250] A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger. Procedural Modeling and Physically Based Rendering for Synthetic Data Generation in Automotive Applications. *arXiv eprint arXiv:1710.06270*, 2017. [page x]
- [251] A. Tsirikoglou, G. Eilertsen, and J. Unger. A Survey of Image Synthesis Methods for Visual Machine Learning. *Computer Graphics Forum*, 39(6):426–451, 2020. [page ix]
- [252] A. Tsirikoglou, K. Stacke, G. Eilertsen, M. Lindvall, and J. Unger. A Study of Deep Learning Colon Cancer Detection in Limited Data Access Scenarios. In *ICLR Workshop on AI for Overcoming Global Disparities in Cancer Care (AI4CC)*. arXiv, 2020. [page x]
- [253] A. Tsirikoglou, M. Gladh, D. Sahlin, G. Eilertsen, and J. Unger. Generative inter-class transformations for imbalanced data weather classification. *London Imaging Meeting*, 2021(1):16–20, 2021. [page xi]
- [254] A. Tsirikoglou, K. Stacke, G. Eilertsen, and J. Unger. Primary Tumor and Inter-Organ Augmentations for Supervised Lymph Node Colon Adenocarcinoma Metastasis Detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 624–633. Springer International Publishing, 2021. [page xi]
- [255] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [page 59]
- [256] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. [page 70]
- [257] B. Uria, I. Murray, and H. Larochelle. RNADE: The Real-Valued Neural Autoregressive Density-Estimator. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2175–2183, Red Hook, NY, USA, 2013. Curran Associates Inc. [page 40]
- [258] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel Recurrent Neural Networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48 of ICML'16, pages 1747–1756. JMLR.org, 2016. [pages 6 and 40]
- [259] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning From Synthetic Humans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [pages 24 and 58]

- [260] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to Unseen Domains via Adversarial Data Augmentation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. [page 70]
- [261] H. Wang, C. Xiao, J. Kossaifi, Z. Yu, A. Anandkumar, and Z. Wang. AugMax: Adversarial Composition of Random Augmentations for Robust Training. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. [page 70]
- [262] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-Video Synthesis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. [page 60]
- [263] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. [pages 46 and 59]
- [264] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. [page 59]
- [265] Y. Wang, L. Zhang, and J. Van De Weijer. Ensembles of Generative Adversarial Networks. *NIPS 2016 Workshop on Adversarial Training*, 2016. [page 65]
- [266] Z. Wang and F. Lu. Single Image Intrinsic Decomposition with Discriminative Feature Encoding. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019. [pages 24 and 58]
- [267] Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402, 2003. [page 44]
- [268] Z. Wang, Q. She, and T. E. Ward. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. *ACM Comput. Surv.*, 54(2), 2021. [pages 39 and 45]
- [269] G. J. Ward. Measuring and Modeling Anisotropic Reflection. *SIGGRAPH Comput. Graph.*, 26(2):265272, 1992. [page 29]
- [270] B. Weyrauch, B. Heisele, J. Huang, and V. Blanz. Component-Based Face Recognition with 3D Morphable Models. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 85–85. IEEE, 2004. [page 24]
- [271] T. Whitted. An Improved Illumination Model for Shaded Display. *Commun. ACM*, 23(6):343–349, 1980. [page 37]

- [272] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing Medical Imaging Data for Machine Learning. *Radiology*, 295(1):4–15, 2020. [page 2]
- [273] J. S. Winter and E. Davidson. Governance of artificial intelligence and personal health information. *Digital Policy, Regulation and Governance*, 21(3):280–290, 2019. [page 53]
- [274] M. Wrenninge and J. Unger. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. *arXiv preprint arXiv:1810.08705*, 2018. [pages 24, 31, 38, and 57]
- [275] H. Wu, S. Zheng, J. Zhang, and K. Huang. GP-GAN: Towards Realistic High-Resolution Image Blending. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, page 24872495, New York, NY, USA, 2019. Association for Computing Machinery. [page 59]
- [276] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876*, 7(8), 2015. [page 68]
- [277] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le. Adversarial Examples Improve Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [page 70]
- [278] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018. [page 44]
- [279] Y. Xue, Q. Zhou, J. Ye, L. R. Long, S. Antani, C. Cornwell, Z. Xue, and X. Huang. Synthetic Augmentation and Feature-Based Filtering for Improved Cervical Histopathology Image Classification. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 387–396, Cham, 2019. Springer International Publishing. [page 71]
- [280] Y. Xue, J. Ye, Q. Zhou, L. R. Long, S. Antani, Z. Xue, C. Cornwell, R. Zaino, K. C. Cheng, and X. Huang. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. *Medical Image Analysis*, 67:101816, 2021. [page 71]
- [281] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6882–6890, 2017. [page 59]
- [282] Z. Yi, H. Zhang, P. Tan, and M. Gong. DualGAN: Unsupervised Dual Learning for Image-To-Image Translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [pages 47 and 59]

- [283] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016. [pages 62 and 63]
- [284] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [page 60]
- [285] X. Yue, B. Wu, S. A. Seshia, K. Keutzer, and A. L. Sangiovanni-Vincentelli. A LiDAR Point Cloud Generator: From a Virtual World to Autonomous Driving. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR '18*, pages 458–464. ACM, 2018. [page 24]
- [286] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. [page 69]
- [287] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. [page 47]
- [288] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916, 2017. [page 59]
- [289] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018. [pages 69 and 70]
- [290] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-Based Rendering for Indoor Scene Understanding Using Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [page 24]
- [291] Y. Zhang, W. Qiu, Q. Chen, X. Hu, and A. Yuille. UnrealStereo: Controlling Hazardous Factors to Analyze Stereo Vision. In *2018 International Conference on 3D Vision (3DV)*, pages 228–237. IEEE, 2018. [page 24]
- [292] L. Zhao, T. Liu, X. Peng, and D. Metaxas. Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Robustness. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14435–14447. Curran Associates, Inc., 2020. [page 70]
- [293] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random Erasing Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, 2020. [page 69]

- [294] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. [pages 10, 45, 46, 59, and 60]
- [295] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin. Emotion Classification with Data Augmentation Using Generative Adversarial Networks. In D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 349–360, Cham, 2018. Springer International Publishing. [page 71]
- [296] C. Zimmermann and T. Brox. Learning to Estimate 3D Hand Pose From Single RGB Images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. [page 24]

Papers

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

<https://doi.org/10.3384/9789179291754>

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2202, 2022
Department of Science and Technology

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se

li.u LINKÖPING
UNIVERSITY