

Dissertation

Machine Learning for Biomedical Applications: From Crowdsourcing to Deep Learning

Shadi N. M. Albarqouni





Technische Universität München

Fakultät für Informatik

Lehrstuhl für Informatikanwendungen in der Medizin

Machine Learning for Biomedical Applications: From Crowdsourcing to Deep Learning

Shadi N. M. Albarqouni

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r): Prof. Dr.-Ing. Darius Burschka

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab

2. Prof. Dr. Dan Stoyanov

University College London, United Kingdom

Die Dissertation wurde am 18.07.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 01.11.2017 angenommen.

Shadi N. M. Albarqouni

Machine Learning for Biomedical Applications: From Crowdsourcing to Deep Learning

Dissertation, Version 0.1

Technische Universität München

Fakultät für Informatik

Lehrstuhl für Informatikanwendungen in der Medizin

Boltzmannstraße 3

85748 and Garching bei München

Abstract

Machine learning is a technique that fosters many Artificial Intelligence Applications in both Computer Vision and Medical Imaging. However, applying this technique blindly, in particular for medical applications, might lead to undesirable performance. Therefore, one must be aware of possible pitfalls, and associated challenges present in machine learning phases, including pre-processing, learning, and evaluation. This calls for incorporating domain-specific knowledge, which is extremely important and plays a crucial role in these applications, into the learning process. In this thesis, a set of mathematical and technical methods incorporating this knowledge is introduced in both unsupervised and supervised setups for different biomedical applications, namely Breast Cancer Histology Imaging, Cryo-Electron Tomography, and Depth Perception in Interventional Imaging.

In Breast Cancer Histology Images, where data can be easily acquired in clinical routines, obtaining such a ground-truth label is a tedious, time-consuming, and rather challenging task for physicians, in particular when the intra-variability agreement between physicians is pretty low. Crowdsourcing is considered the state-of-the-art for collecting inexpensive image annotations. However, it is still hampered by the need of domain knowledge and expertise for medical data. In this context, a robust aggregation layer for a deep learning framework, which combines both Human and Artificial intelligence, is proposed to generate ground-truth annotations from noisy annotations collected by crowdsourcing, as well as play-sourcing platforms.

However, in some cases, due to the absence of data abundance and labels, *e.g.* in Cryo-Electron Tomography, regularization based methods that incorporate prior knowledge, for example, Huber terms for robust regression, and Graph Laplacians for manifold embedding, are proposed in the context of tomographic reconstruction and noise reduction, respectively. In Interventional Imaging, prior knowledge information obtained from pre-operative Computed Tomography (CT) scans are used to obtain labels for machine learning algorithms, to improve the visual perception of interventional X-ray images.

Finally, this thesis concludes incorporating such a prior knowledge, *i.e.* domain-specific knowledge, intelligently in the learning phase, influencing the performance positively.

Keywords: Regularization, Laplacian Graph, Dictionary Learning, Deep Learning, Crowdsourcing, Gamification.

Zusammenfassung

Maschinelles Lernen ist eine Technik, die viele Künstliche Intelligenz Anwendungen, sowohl in Computer Vision als auch in Medical Imaging fördert. Allerdings könnte die gedankenlose Anwendung dieser Technik, insbesondere für medizinische Anwendungen, zu einer unerwünschten Leistung führen. Daher muss man sich über mögliche Fallern und damit verbundene Herausforderungen in maschinellen Lernphasen, einschließlich Vorverarbeitung, Lernen und Auswertung, bewusst sein. Dies erfordert die Einbindung von domänenspezifischem Wissen, das äußerst wichtig ist und bei diesen Anwendungen eine entscheidende Rolle spielt, in den Lernprozess. In dieser Arbeit wird eine Reihe von mathematischen und technischen Methoden, die dieses Wissen einbeziehen, sowohl in unbeaufsichtigter als auch in beaufsichtigter Konfiguration für verschiedene biomedizinische Anwendungen eingeführt, nämlich histologische Bildverarbeitung für Brustkrebsanalyse, Cryo-Elektronen Tomographie und visuelle Wahrnehmung in interventioneller Bildgebung.

In den Brustkrebs-Histologie-Bildern, wo Daten in klinischen Routinen leicht erworben werden können, ist jedoch die Erlangung eines solchen annotierten Vorwissens eine langwierige, zeitaufwändige und anspruchsvolle Aufgabe für Ärzte, insbesondere wenn die intravariablen Vereinbarung zwischen Ärzten sehr niedrig ist. Crowdsourcing gilt als aktueller Stand der Technik für das Sammeln von preiswerten Bild-Annotationen. Allerdings ist diese Methode immer noch eingeschränkt durch die Notwendigkeit von Fachkenntnissen und Know-how insbesondere für medizinische Daten. In diesem Zusammenhang wird eine robuste Aggregationsschicht für ein Deep Learning Framework vorgeschlagen, das sowohl menschliche als auch künstliche Intelligenz kombiniert, um genauere Annotationen aus fehlerbehafteten Annotationen zu generieren, die durch Crowdsourcing sowie Play-Sourcing-Plattformen gesammelt wurden.

In einigen Fällen werden jedoch aufgrund der Abwesenheit von Datenreichtum und annotierten Daten, z. B. in der Cryo-Elektronen-Tomographie, regularisationsbasierte Methoden, die Vorkenntnisse beinhalten, beispielsweise Huber-Begriffe für eine robuste Regression und Graph Laplace-Operatoren für vielfältige Einbettung im Rahmen der tomographischen Rekonstruktion und Rauschreduktion, vorgeschlagen. In der interventionellen Bildgebung werden vorherige Kenntnisse, die aus präoperativen Computer Tomographie (CT) Bildern gewonnen wurden, verwendet, um Annotationen für maschinelle Lernalgorithmen zu erhalten, um die visuelle Wahrnehmung von interventionellen Röntgenbildern zu verbessern.

Letztendlich beschreibt diese Doktorarbeit die Einbeziehung solcher Vorkenntnisse, d.h. Domänenspezifisches Wissen auf intelligente Art und Weise in die Lernphase, wodurch die Leistung positiv beeinflusst wird.

Keywords: Regularization, Laplacian Graph, Dictionary Learning, Deep Learning, Crowdsourcing, Gamification.

Acknowledgments

First of all, I would like to express my sincere gratitude to Prof. Dr. Nassir Navab for giving me the opportunity to do my research in collaboration with the Computer Aided Medical Procedures (CAMP) group, at Technical University of Munich (TUM) in Germany and Johns Hopkins University (JHU) in USA. Prof. Navab has provided me an endless support and encouragement throughout the last four years. Without that, this thesis would not have been possible.

I would also like to appreciate all contributions from my colleagues during my doctoral study, in particular, the senior co-authors; Tobias Lasser, Maximilian Baust, and Stefanie Demirci, and this gratitude also extends to my colleagues; Diana Mateus, Lichao Wang, Tingying Ping, Sailesh Conjeti, Mohammed Alsheikhali, Wadim Kehl, Loic Peter, and Martina Hilla for their help, support, and fruitful discussion. Furthermore, I would like to thank Ashraf Al-Amoudi, and Weaam Alkhalidi from Deutsches Zentrum für Neurodegenerative Erkrankungen e. V. (DZNE) for their help and support, in particular, at the beginning of my Ph.D. journey.

Finally, I would like to say thank you to my parents; Sabah and Nabil, my sisters; Shatha, Roba, Aya, and Dima, and my brothers; Loai and Mohammed for their continuous support over the years. Last but not least, I would like to thank my wife Reem and my daughter Alma for their patience and tremendous support.

Contents

1	Introduction	3
1.1	Motivation	3
1.2	Problem Statement	5
1.3	Roadmap	6
2	Towards Deep Learning for Medical Applications	7
2.1	Pre-Processing	7
2.1.1	Feature normalization	7
2.1.2	Missing Data	9
2.1.3	Class Imbalance	9
2.2	Learning	10
2.2.1	Logistic Regression	14
2.2.2	Linear Regression	14
2.2.3	Dictionary Learning	16
2.2.4	Deep Learning	17
2.3	Evaluation	18
2.3.1	Classification Metrics	19
2.3.2	Image Quality Metrics	21
3	From Nano- to Milli-meter Image Resolution	23
3.1	Cryo-Electron Tomography	23
3.1.1	Problem Definition and Motivation	24
3.1.2	Related Work	25
3.1.3	Contribution: Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction (CMMI MICCAI 2014)	27
3.1.4	Contribution: Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data (BMVC 2015)	27
3.2	Breast Cancer Histology Images	28
3.2.1	Problem Definition and Motivation	28
3.2.2	Related Work	30
3.2.3	Contribution: AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images (IEEE TMI 2016)	30
3.2.4	Contribution: Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research (LABELS/DLMIA MICCAI 2016)	31
3.3	Depth Perception in Interventional X-ray Imaging	32
3.3.1	Related Work	33
3.3.2	Contribution: Single-view X-ray depth recovery: toward a novel concept for image-guided interventions (IJCARs 2016)	33

3.3.3	Contribution: X-ray In-Depth Decomposition: Revealing The Latent Structures (MICCAI 2017)	34
4	Conclusion and Outlook	35
	List of Figures	37
	List of Tables	41
	Bibliography	43
A	Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction	59
B	Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data	61
B.1	Introduction	62
B.2	Methodology	63
B.2.1	Graph Representation	63
B.2.2	Graph Spectral Filtering	64
B.2.3	Connection to Classical Filters	65
B.2.4	Stopping Criterion	65
B.3	Experiments and Results	65
B.4	Conclusion	68
B.5	References	68
C	AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images	71
C.1	Introduction	72
C.2	Methodology	74
C.2.1	Notation	75
C.2.2	Multi-scale CNN Model	75
C.2.3	Aggregation Layer (AG):	75
C.3	Experiments and Results	77
C.3.1	Proof-of-Concept Evaluation	79
C.3.2	Use Case Evaluation	84
C.4	Discussion	85
C.5	Conclusion	87
C.6	References	88
D	Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research	91
E	Single-view X-ray depth recovery: toward a novel concept for image-guided interventions	93
F	X-ray In-Depth Decomposition: Revealing The Latent Structures	95
G	Abstracts of Publications not Discussed in this Thesis	97

List of Authored and Co-authored Publications

Discussed in This Dissertation

- [1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. “AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1313–1321.
- [2] S. Albarqouni, T. Lasser, W. Alkhaldi, A. Al-Amoudi, and N. Navab. “Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction”. In: *Computational Methods for Molecular Imaging*. 2015, pp. 43–51.
- [3] S. Albarqouni, M. Baust, S. Conjeti, A. Al-Amoudi, and N. Navab. “Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data.” In: *British Machine Vision Conference (BMVC)*. 2015, pp. 17–1.
- [4] S. Albarqouni, S. Matl, M. Baust, N. Navab, and S. Demirci. “Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research.” In: *LABELS/DLMIA@ MICCAI*. 2016, pp. 269–277.
- [5] S. Albarqouni, U. Konrad, L. Wang, N. Navab, and S. Demirci. “Single-view X-ray depth recovery: toward a novel concept for image-guided interventions”. In: *International journal of computer assisted radiology and surgery* (2016), pp. 1–8.
- [6] S. Albarqouni, J. Fotouhi, and N. Navab. “X-ray In-Depth Decomposition: Can Deep Learning Reveal The Latent Structures?” In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017*. Vol. Lecture Notes in Computer Science Volume 10435. 2017, pp. 401–409.

Selected Publications

- [7] C. Baur, S. Albarqouni, S. Demirci, N. Navab, and P. Fallavollita. “CathNets: Detection and Single-View Depth Prediction of Catheter Electrodes”. In: *International Conference on Medical Imaging and Virtual Reality*. Springer. 2016, pp. 38–49 (cit. on p. 36).
- [8] C. Baur, S. Albarqouni, and N. Navab. “Semi-Supervised Learning for Fully Convolutional Networks”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017*. Vol. Lecture Notes in Computer Science Volume 10435. 2017, pp. 281–289 (cit. on p. 36).

- [9] M. Bui, S. Albarqouni, M. Schrapp, N. Navab, and S. Ilic. “X-Ray PoseNet: 6 DoF Pose Estimation for Mobile X-Ray Devices”. In: *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE. 2017, pp. 1036–1044 (cit. on p. 36).
- [10] A. Vahadane, T. Peng, S. Albarqouni, et al. “Structure-preserved color normalization for histological images”. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE. 2015, pp. 1012–1015 (cit. on p. 29).
- [11] A. Vahadane, T. Peng, A. Sethi, et al. “Structure-preserving color normalization and sparse stain separation for histological images”. In: *IEEE transactions on medical imaging* 35.8 (2016), pp. 1962–1971 (cit. on p. 29).

Introduction

“ *If learning the truth is the scientist’s goal... then he must make himself the enemy of all that he reads.*

— **Al-Hazen Ibn Al-Haytham (965-1040CE)**

A Brief History: The brilliant breakthrough in both vision and light took place in the 9th century when Alhazen Ibn Al-Haytham invented the *Camera Obscura* (in Latin) or *Albait Almuzlim* (in Arabic), which simply means the dark room. Ibn Al-Haytham observed that the act of vision is accomplished by rays emitting from external objects and entering the visual organs, demolishing the extra-mission theories of his predecessors. Ibn Al-Haytham reported his systematic experiments and theories in his great work, *Kitab Al-Manazir* or *Book of Optics* [18], which has been ranked as one of the most influential books ever written in the history of optics. His ideas influenced many Western scholars enabling them to develop optical microscopy in the 14th and 15th centuries until Isaac Newton wrote his "new" theory about light and colors in 17th century [182]. In his work, he identified the colors which form the visible band within the electromagnetic spectrum. Afterward, electromagnetic radiations other than visible lights were discovered including infrared radiation, ultraviolet radiation, and radio waves. In the late 19th century, X-ray was first discovered by Roentgen who noticed that rays were able to penetrate soft tissues, *i.e.* Human flesh better than hard ones, *i.e.* bones. This discovery enabled both diagnostic and interventional imaging such as mammography, Computed Tomography (CT), Fluoroscopy and Angiography.

1.1 Motivation

With the rapid and advanced development in information and communication technology (ICT), the amount of data is increasing exponentially. For example, the healthcare system in the United States alone produced more than 150 exabytes (150×10^{18}) back in 2011 [198] and it is expected that the worldwide health records will reach 40 yottabytes (40×10^{24}) in 2020 [181]. This big data, as defined in [168] by

“the ability of society to harness information in novel ways to produce useful insights or goods and services of significant value”,

is making a difference in public health sector, *i.e.* predicting outbreaks, improving patient care and health outcomes [168, 198]. However, this numerous application of big data can not happen without statistical and automated models for data analysis; pattern recognition, prediction, decision making, that govern the learning system, which is what Artificial Intelligence (AI), and Machine Learning (ML) provide.

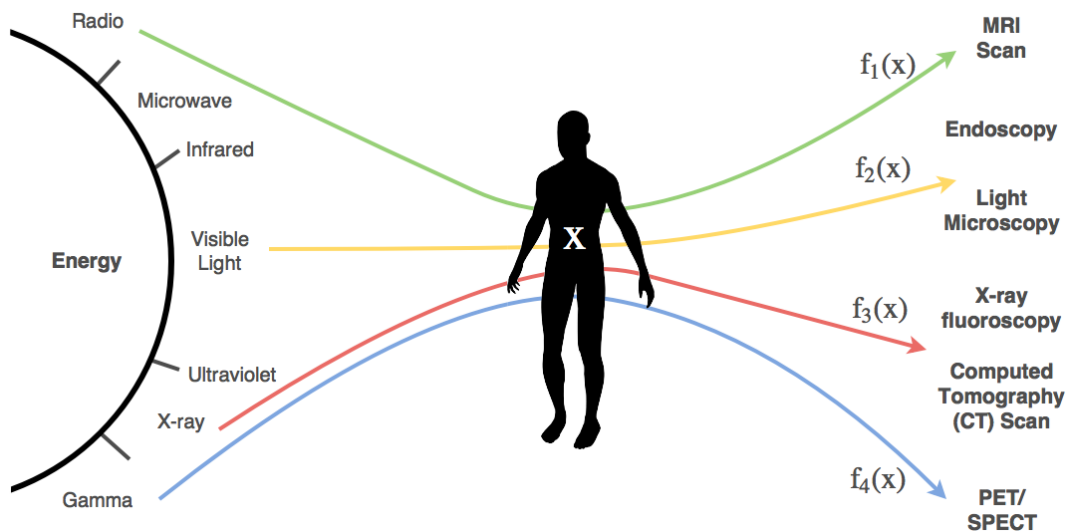


Figure 1.1. Energy sources and different imaging modalities.

Over the last decade, most research in machine learning has emphasized the use of handcrafted features, and data-driven based approaches to solving different tasks in the medical field (*cf.* Fig.1.2). Recently, with the advances in graphic processing units (GPU), computing, and optimization, deep learning has emerged as a powerful tool enabling us to solve challenging tasks in both Computer Vision [133] and Medical Imaging [70]. One of the most significant applications of AI and ML in the medical field is Computer Aided Diagnosis (CAD) [58] that aims at assisting clinicians and physicians based on different information sources such as Electronic Medical Records (EMR), and different imaging modalities, *i.e.* X-ray, CT, and Magnetic Resonance Imaging (MRI).

However, as pointed out by Andrew Ng [183], Artificial Intelligence (AI) and Machine Learning (ML), in general, can not work without a **huge amount of data** and **talent**, referring to *domain-specific knowledge*. Similarly, Ge Wang [255] holds the view that developing a new generation of image reconstruction techniques might lead to elegant utilization of *domain-specific knowledge* as prior knowledge, and consequently boosting the performance in several clinical applications.

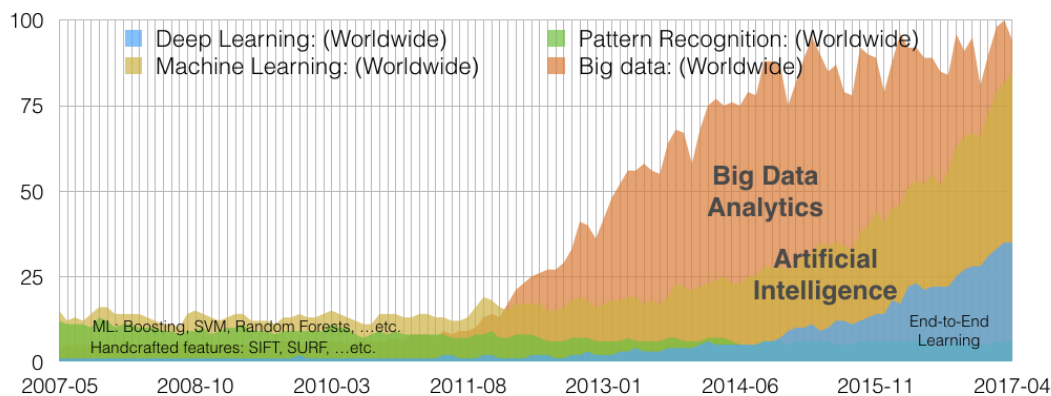


Figure 1.2. Google trends on Big Data and Artificial Intelligence; Pattern Recognition, Machine Learning, and Deep Learning.

1.2 Problem Statement

In this dissertation, we address the following questions, which are considered important for machine learning problems, in particular for medical applications, covering the pre-processing, learning, and evaluation phases, respectively:

- **To what extent can Artificial Intelligence (AI) and Human Intelligence (HI) collaborate for robust Biomedical Image Annotation?**

Given the lack of publicly available ground-truth, transferring recent developments in deep learning to the biomedical imaging becomes a serious challenge. Crowd-sourcing has enabled us to collect annotations for large scale databases, *i.e.* ImageNet¹. However, its application to biomedical imaging raises many questions about the reliability of collected annotations and the agreement between different users. Therefore, combining both AI and HI in a single framework that can generate reliable image annotations is highly desirable.

- **How can prior knowledge be defined and incorporated into Machine Learning (ML) algorithms?**

Employing machine learning algorithms blindly, particularly in medical applications, without deeply understanding the challenging circumstances, might lead to unsatisfactory performance. This urges the demand for a domain-specific knowledge that can be incorporated into machine learning algorithms, starting from simple feature extraction methods, *i.e.* handcrafted engineered features, to more sophisticated regularized energy functions. Insights of the definition and formulation of such a prior knowledge are expected to be delivered for different biomedical applications and machine learning tools as well.

- **What are the challenges present in real scenarios, how this would affect the evaluation, and what are the possible solutions?**

A limited amount of labeled data, highly imbalanced classes, different scanners, together with inter- and intra-observer variability are common challenges in medical applications. For instance, training an ML algorithm from highly imbalanced data, where the majority of samples are negative, and few samples are positive, would negatively influence the performance. This requires new techniques handling the under-represented classes, *i.e.* re-sampling, and data augmentation, and independent metrics that can measure the performance regardless the class distribution.

The challenges above are essential for the study of machine learning for biomedical applications, and we believe that our novel contributions, presented in this dissertation, provide answers covering many challenging aspects, *from Nano- to Milli-meter image resolution, and from Crowdsourcing to Deep Learning* approaches.

¹<http://www.image-net.org/>

1.3 Roadmap

This dissertation is structured as follows:

In Chapter 2, we briefly define machine learning (ML) and the relevant technical questions in different ML phases. The chapter is divided into three main sections, in each, we describe ML phase along with its corresponding challenges. For instance, preprocessing steps along with the techniques handling missing data and class imbalance are discussed in Sec. 2.1. Afterward, Sec. 2.2 focuses on the learning phase of ML, *i.e.* Big Data Vs. Small Data, Bias-Variance problem, before we explain the mathematical tools used in this dissertation. Evaluation metrics in the presence of imbalanced data are discussed in Sec. 2.3.

Moving forward, Chapter 3 describes the ML application on different imaging modalities, starting *from Nano- to Millimeter Image Resolution*. In sec. 3.1, we introduce the Cryo-Electron Tomography and its impact on understanding and interpretation of complex biological structures. Then we present the major challenges associated with the electron microscopy imaging device, *i.e.* limited angle view, radiation damage, and the contrast transfer function before we present our contributions in both 3D Tomographic Reconstruction and Noise Reduction. Moving to the micro-scale (*cf.* Sec. 3.2), we briefly introduce the Breast Cancer Histology Images and the clinical procedures, as well as the challenges associated with the staining and observer disagreement. Insights about our participation in AMIDA13 Challenge, related work, and our significant contributions in Crowdsourcing and Gamification are also presented. After that, Sec. 3.3 highlights the need for Depth Perception in Interventional X-ray imaging and presents our most significant contributions, *i.e.* Depth recovery of single-view X-ray images, and X-ray In-Depth Decomposition.

Finally, Chapter 4 combines our conclusion and future outlook by discussing the importance of prior knowledge to various biomedical applications, and further list open questions that could be food for thought for any future research.

Towards Deep Learning for Medical Applications

“*The good teacher makes the poor student good and the good student superior!*”

— Marva Collins (1936–2015)

As defined by Jordan *et al.* [124], ML is

“The discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience?, and What are the fundamental statistical computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?”.

It is lying at the intersection of computer science and mathematics & statistics as represented in the Venn diagram of Drew Conway (*cf.* Fig.2.1). Nowadays, with the advent of big data, scalable machine learning techniques are demanded, namely deep learning, which has been recently emerged as a powerful tool handling this massive amount of data (*cf.* Fig.2.1).

In 2017, Zhou *et al.* [288] proposed a framework for ML on big data (MLBiD), which interacts mainly with four components, *i.e.* *big data*, *user*, *domain knowledge*, and *system*. Opportunities and challenges with respect to aforementioned components are thoroughly discussed. As depicted in Fig.2.2, all interactions are bi-directional, for instance; *big data* acts as input to ML algorithm outputting a useful information that sent back to form a big data, *user* interacts with ML providing the domain knowledge, usability, and design requirements, and in return ML assists users in decision making, *domain* acts as a source of knowledge and the area where ML can be applied as well; *system* determines the running time of ML algorithms, where ML can define the system design requirements.

This chapter is not intended to provide a comprehensive overview of machine learning but rather focus on common technical questions in different ML phases; Pre-processing, Learning, and Evaluation. Readers are referred to [27] to understand the basic concepts of machine learning algorithms.

2.1 Pre-Processing

2.1.1 Feature normalization

It has been shown in the literature that this process is very essential for many machine learning algorithms. For instance, having a high-dimensional feature vector $\mathbf{x} \in \mathbb{R}^n$ that has some attributes in range (0, 1000), and others in range $(-1, 1)$, would highly bias the learning process. To rescale the

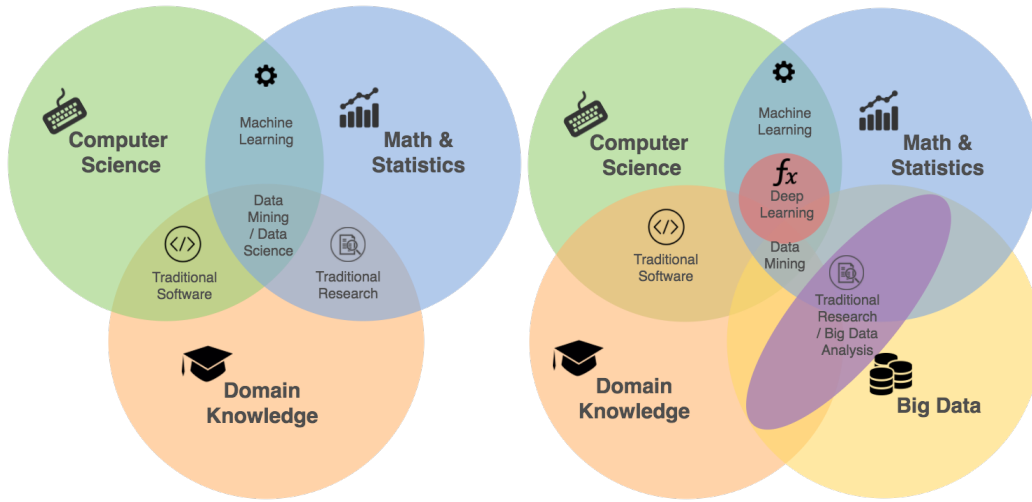


Figure 2.1. Venn Diagram of Data Science: Drew Conway [218] (left), Modified one (right).

dynamic range, commonly used normalization and standardization techniques are presented for a given feature matrix $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{n \times N}$, where n is the number of instances.

Normalization. Given a lower bound $\mathbf{l}_x \in \mathbb{R}^n$ and an upper bound $\mathbf{u}_x \in \mathbb{R}^n$, the normalized feature matrix \bar{X} is rescaled into range $[0,1]$ as

$$\bar{X} = \frac{X - \mathbf{l}_x}{\mathbf{u}_x - \mathbf{l}_x}, \quad (2.1)$$

Standardization. Given a mean $\mu_x \in \mathbb{R}^n$ and a covariance matrix $\Sigma_x \in \mathbb{R}^{n \times n}$, the standardized feature matrix \bar{X} is transformed to a random variable with zero mean and unit covariance as

$$\bar{X} = \frac{X - \mu_x}{\Sigma_x}, \quad (2.2)$$

where $\Sigma_x = \frac{1}{N-1} \sum_{i=1}^N (X - \mu_x)(X - \mu_x)^T$. Other rescaling techniques such as Whitening, Rank normalization, and Zero Component Analysis (ZCA), are widely covered in the literature [14, 132].

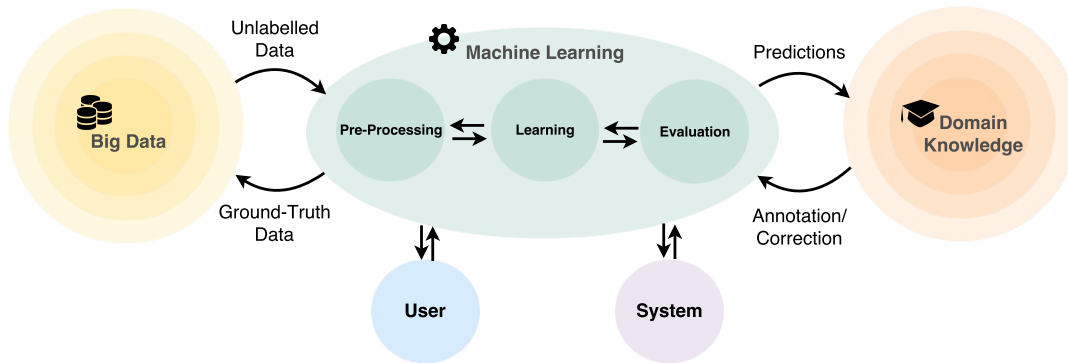


Figure 2.2. MLBiD framework. Adapted from Fig.1 in [288].

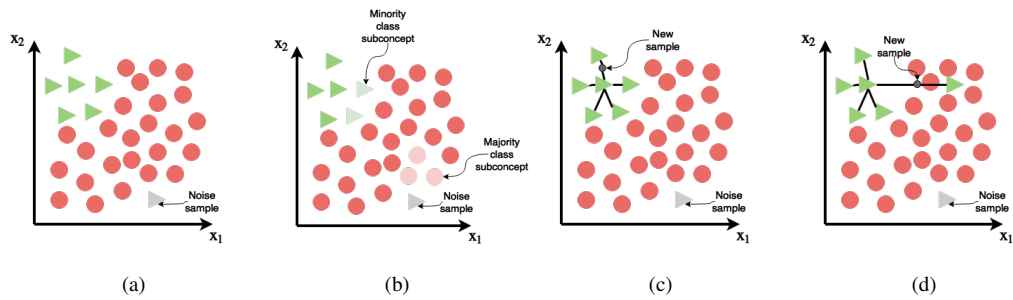


Figure 2.3. Class Imbalance in (a) classical scenarios, and (b) complex ones with the presence of intra- and inter-class imbalance. Example on synthetic methods, and its drawback in (c) and (d).

2.1.2 Missing Data

In real applications, high-dimensional feature vectors often associated with missing and incomplete data. Replacing the missing data with substituted values is so-called Data Imputation. Simple techniques such as mean substitution, and hot-deck, are known to increase the risk of model bias. Therefore, sophisticated learning-based methods are proposed showing a significant performance over the statistical-based ones [121].

2.1.3 Class Imbalance

Learning from highly imbalanced data, where few positive samples are available among the majority of negative samples (*cf.* Fig. 2.3), highlighted as a major challenge yielding a serious problem, *i.e.* model bias, in ML research. To mitigate this issue, few methods are presented here, while a comprehensive review of different methods is covered in [112].

Sampling Methods While simple random over-sampling or under-sampling techniques can be performed to balance the class distribution, it so happened that worsen the performance. For instance, under-sampling the majority class could lead to missing important features or patterns. In contrast, over-sampling the minority class might lead to *overfitting* (*cf.* Sec. 2.2) due to so many replicated samples. One powerful technique, namely SMOTE method [40], has shown great success in many applications. It creates basically synthetic instances that are close to the minority class.

Given an instance \mathbf{x}_i in the minority class, and its K -nearest neighbors (k -NN), the synthesized new instance \mathbf{x}_{new} can be obtained using one randomly sampled of k -NN \mathbf{x}_k as

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_i - \mathbf{x}_k), \quad (2.3)$$

where λ is a random number between $[0,1]$.

One drawback of synthetic sampling methods is the possible overlapping between the synthetic example and the majority class (*cf.* Fig. 2.3). Data cleaner methods such as Tomek Link [236] are used to clear unwanted overlapping samples.

Cost Sensitive Methods. Alternatively, penalties associated with misclassifying examples are considered in the cost-sensitive methods. For example, in a binary classification scenario, a weighting scheme can be used in cross entropy loss function as

$$\mathcal{L}(\hat{y}, y) = \sum_{i=1}^N \alpha \cdot y_i \log \hat{y}_i + (1 - \alpha) \cdot (1 - y_i) \log(1 - \hat{y}_i), \quad (2.4)$$

where \hat{y}_i is the predicted value, y_i is the corresponding ground-truth, and α is the cost penalty on misclassifying positive class. Readers are referred to Sec. 2.2.1 for more details about logistic regression for classification tasks.

2.2 Learning

How much data we need for a machine learning algorithm?

Big Data vs. Small Data. Variability and heterogeneity available in medical data calls for personalized medicine that provides a proper decision for an individual patient for the specific medical procedure. Therefore, in contrast to the potential impact of big data in healthcare, Sacristan *et al.* [212] spotted some potential barriers that may arise limiting the development of big data in research and medicine, in particular, the knowledge transfer from clinical researches (Big data) to medical care (small data) and vice versa. They believed that there is “no big data without small data” and policy-makers should care about the quality of small data, *i.e.* completeness, standardization. Therefore, such a learning healthcare system that proposed in [212] is desirable, where the medical act (doctor-patient encounter) plays a crucial role in the learning process (*cf.* Fig. 2.4). In another study, Ferguson *et al.* [72] have shown in their paper, entitled “big data from small data”, that small data which carries a lot of diversity and heterogeneity might not be suitable for publications. However, a collection of such small datasets, so-called long-tail data, can be utilized and in fact turned into a big data (*cf.* Fig. 2.4).

Léon Bottou, from Facebook AI Research, pointed out in his presentation on Big Data [28] that the assumption of data scarcity advanced the development of statistical machine learning, and having Big data nowadays should not be utilized to increase the average accuracy rather mitigate the challenging aspects, in reality, *i.e.* Transfer Learning, and Domain Adaptation.

In medical field, having such a standardized or complete data is a bit challenging for many reasons. For instance, biomedical data acquisition takes from few minutes for MR, or CT scans to a couple of weeks for Cryo-Electron Tomography [108] and needs a lot of preprocessing steps including image reconstruction, noise-reduction, registration, and most importantly labeling or annotation. Furthermore, medical data need to be de-identified and anonymized for privacy issues [83, 136]. All these issues hinder the availability of such a large scale database that can be utilized for ML algorithms in healthcare systems. Having said that, massive efforts are made to collect and gather de-identified, and standardized medical data from multi-research centers across the world, giving an excellent opportunity for scientists and researchers to contribute and evaluate their algorithms on benchmark databases. Table 2.1 lists few challenges, offered at grand-challenges website¹, that provide multi-centers clinical database.

¹Grand-challenges website, https://grand-challenge.org/All_Challenges/

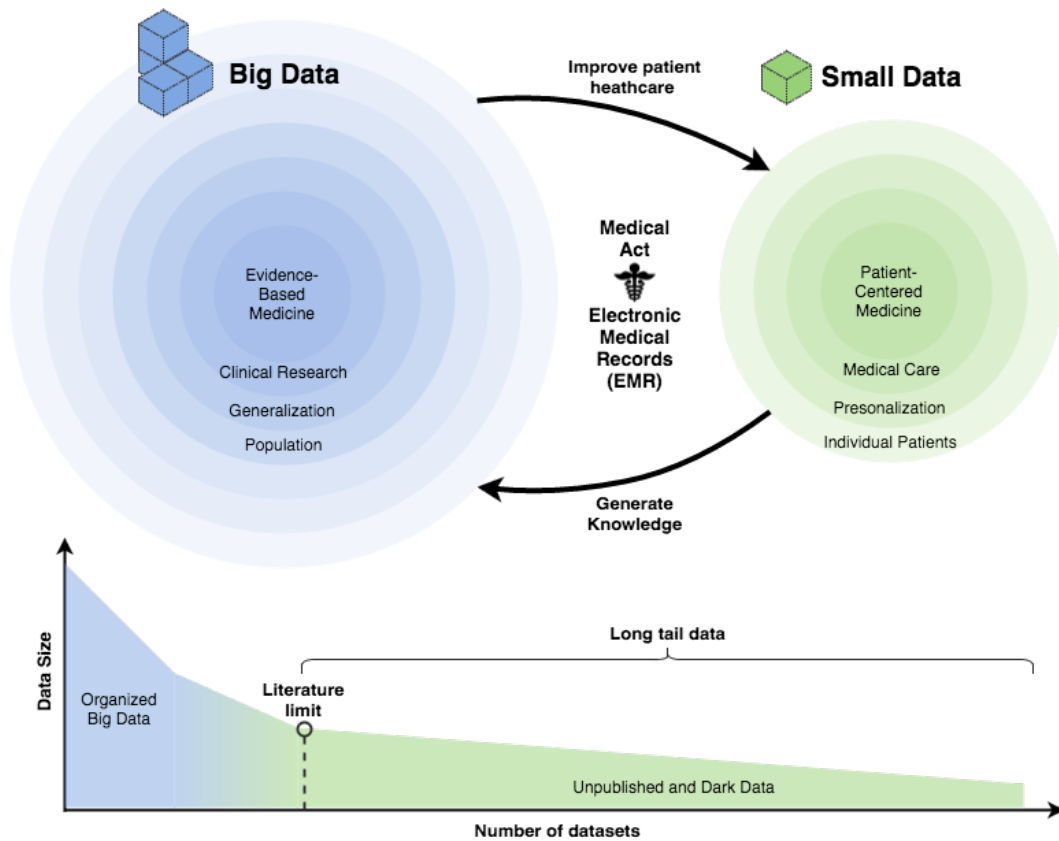


Figure 2.4. Big data or Small data: On top is the learning healthcare system proposed by Sacristan *et al.* [212], while the bottom figure, adapted from [72], shows the long tail of unpublished and dark datasets, which can be a potential treasure for big data.

Given the aforementioned challenges and aspects of having such a big medical database, the question remains, how we can define *enough data* for an ML algorithm?. It is still an open question. However, some technical hints can guide the ML practitioners to understand the impact of using a small amount of data, so-called variance problem.

Bias-Variance tradeoff. It is considered as one of the common problems in Machine Learning. *Bias* error is mainly due to the improper choice of model complexity. For instance, approximating a real-life problem, which might be extremely complicated, by a pretty simple model. This bias results in undesired performance regardless the number of instances. One example is modeling a highly complex function by a simple linear regression. In contrast, the *Variance* error happens due to the limited number of instances in the training set. Ideally, the trained model should generalize well on new instances. However, due to the small number of training sample, the model tends to overfit and produces a significant discrepancy, so-called variance error, for a new instance. Figure 2.5 shows the performance of different models in the presence of bias and variance. That just means, whenever the model complexity fits the problem, you can start with a limited amount of data and keep increasing it gradually until you minimize the variance error between training and validation losses, and hence we reach the *enough* amount of data. The later problem is also referred to the *overfitting* problem,

²<http://amida13.isi.uu.nl/>

³<https://camelyon16.grand-challenge.org>

⁴<https://retouch.grand-challenge.org/details/>

⁵<https://portal.fli-iam.irisa.fr/msseg-challenge/overview>

Table 2.1. Publicly available challenges and the rank obtained by the author’s team.

Challenge	Modality	Organ	Task	Training Set	Testing Set	Rank
AMIDA13 ²	Histology	Breast	Mitosis Detection	(12) 311 HPF	(11) 305 HPF	3/16
CAMELYON16 ³	Histology	Breast	Cancer metastasis detection	(270) WSI	(130) WSI	19/32
RETOUCH ⁴	OCT	Retina	Fluid detection and segmentation	(72) volumes	(30) volumes	N/A
MSSEG ⁵	MRI	Brain	MS Lesion segmentation	(15) volumes	(38) volumes	N/A

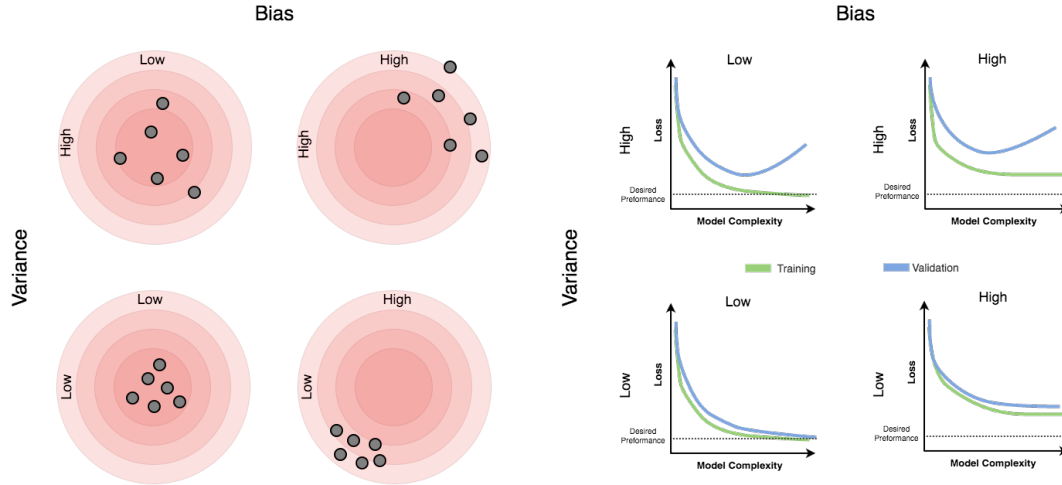


Figure 2.5. Bias-Variance tradeoff: dotted lines refers to the desired performance, where the model complexity refers to the number of iterations in this context.

where the number of instances is much less than the number of parameters (under-determined systems). Regularization techniques are proposed in this context to avoid overfitting and find a *sub-optimal* solution.

What is the difference between different machine learning algorithms?

Before diving into different types of machine learning algorithms, a brief discussion about the two cultures of statistical modeling; **data modeling** and **algorithmic modeling** will be presented. To explain the difference between the two cultures (*cf.* Fig.2.6), we would assume a black box system generates a response variable $\mathbf{y} \in \mathbb{R}^m$ from an input variable $\mathbf{x} \in \mathbb{R}^n$, where the goal is i) to predict the response variable $\hat{\mathbf{y}}$ for any novel input variable, and/or ii) to extract some information about the system $f(\cdot)$. The response variable of a given input \mathbf{x} can be written as

$$\hat{\mathbf{y}} = f(\mathbf{x}; \mathbf{w}), \quad (2.5)$$

where \mathbf{w} are the learned parameters.

Researchers in data modeling culture build some assumptions about the data enabling them to formulate the response as a stochastic model with known distributions, for instance,

$$\hat{\mathbf{y}} = f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^n w_i x_i + \eta, \quad \text{where } \eta \sim \mathcal{N}(0, \sigma), \quad (2.6)$$

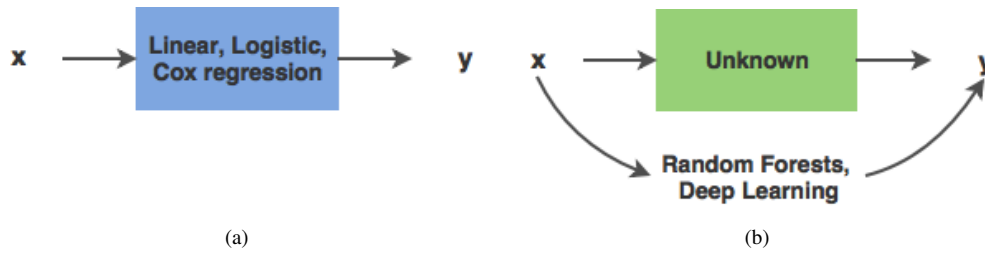


Figure 2.6. Two cultures: a) Data Modeling, b) Algorithmic Modeling.

where $\eta \sim \mathcal{N}(0, \sigma)$ is an *i.i.d* Gaussian noise with zero mean and σ standard deviation, and w are the estimated parameters. Logistic, Linear and Cox Regression are typically used in this community. In contrast, researchers in the algorithmic modeling culture assume $f(x)$ is an unknown complex model that handles *i.i.d* unknown multivariate distribution. Random Forests and Deep Learning are examples of such algorithms (*cf.* Fig. 2.6). Leo Breiman [32] holds the view that the commitment of the statistical community to the data modeling had led to irrelevant theory and questionable conclusion. Further, it has kept the statisticians from working on interesting new problems and using more suitable algorithmic models. His conclusion, which brought a lot of discussions, was drowned from his experience while leading many projects in both academia and industry.

Up to now, we have mentioned few examples on machine learning algorithms where the objective is to predict the response variable given an enough amount of labeled data; so-called supervised learning. However, machine learning algorithms can be further available in different setups such as unsupervised, and semi-supervised learning.

Supervised learning. It refers to ML algorithms where the ground-truth is available in the training set, *i.e.* class labels, or continuous outputs. This ground-truth is pre-assumed that is reliable to govern the learning process; otherwise, it would mislead the learning process.

Unsupervised learning. Unlike the supervised learning, it does not require any ground-truth for the training. However, it mainly depends on the intrinsic structure of the manifold. This setup is commonly used for dimensionality reduction, clustering, and feature representation.

Semi-supervised learning. It refers to ML algorithms where few labeled data are used together with a massive amount of unlabeled data. The objective is to improve the supervised learning task by leveraging this large amount of unlabeled data during the training.

For supervised learning, the general formula of the loss function for a given ground-truth and predicted output,

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \min_f \sum_{i=1}^N \phi(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})), \quad (2.7)$$

where $\phi(\cdot)$ is the underlying loss function that describes the cost of misclassifying (deviating) the ground-truth discrete (continuous) label.

Prior knowledge, *i.e.* domain-specific knowledge, can be incorporated into the loss function, as a soft constraint, to find an optimal or sub-optimal solution. The regularized loss function is formulated using Lagrangian multiplier as

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \min_f \sum_{i=1}^N \underbrace{\phi(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w}))}_{\text{Reconstruction Error}} + \underbrace{\lambda \cdot R(f)}_{\text{Regularization}}, \quad (2.8)$$

where λ is the regularization parameter.

The following sections attempt to give a mathematical background on both cultures starting with the Linear regression, dictionary learning to the recent sophisticated models of deep learning.

2.2.1 Logistic Regression

Logistic Regression is used for classification tasks, where the goal is to assign a given input of an n -dimensional feature space $\mathbf{x} \in \mathbb{R}^n$ to one of K discrete classes C_k , where $k = 1, \dots, K$. To be able to exactly separate the feature space to disjoint classes, feature space is assumed to be linearly separable. The predicted output variable \hat{y} ,

$$\hat{y} = f(\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x} + \omega_0), \quad (2.9)$$

where $\sigma(\cdot)$ is known as an activation function, and its objective to squeeze the continuous values to discrete ones, or more generally to probabilities that lie in the range (0,1). Typical choices for such an activation function are softmax, hyperbolic tangent, and sigmoid functions.

To estimate the learned parameters \mathbf{w} , *cross-entropy* loss function is typically utilized in Eq. 2.7,

$$\phi(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})) = \sum_{k=1}^K -y_{ik} \log \hat{y}_{ik}, \quad (2.10)$$

where \hat{y}_{ik} is the predicted output for the corresponding class k . To avoid overfitting, a smoothness term such as ℓ_2 -norm is added to the loss function similar to Eq. 2.8.

2.2.2 Linear Regression

Linear Regression is similar to the Logistic Regression, however, the response variable $\mathbf{y} \in \mathbb{R}^m$ is continuous. The predicted output variable \hat{y} is expressed as

$$\hat{y} = f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + \omega_0, \quad (2.11)$$

To estimate the learned parameters \mathbf{w} , *Mean-Square-Error* loss function is typically utilized in Eq. 2.7,

$$\phi(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})) = \frac{1}{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2, \quad (2.12)$$

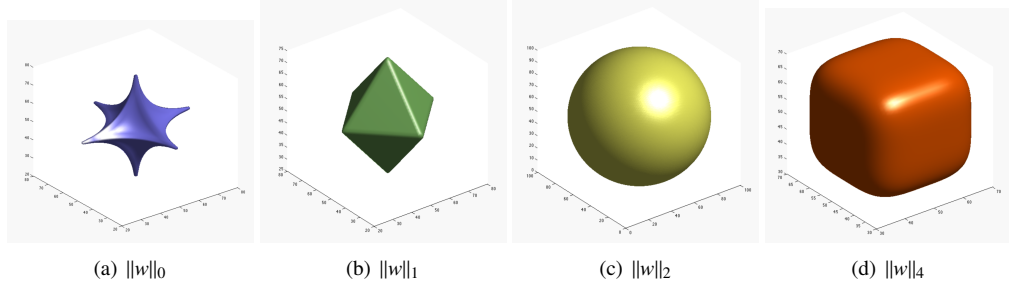


Figure 2.7. The geometry of different norms in 3-dimensional space: Sparsity prior in (a-b), and Smoothness prior in (c-d).

where $\|\cdot\|_2$ is the Euclidean distance. To incorporate a prior knowledge, an ℓ_p -norm is added to the loss function similar to Eq. 2.8.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \min_f \sum_{i=1}^N \phi(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})) + \lambda \|\mathbf{w}\|_p, \quad (2.13)$$

where $\|\cdot\|_p$ is the ℓ_p -norm in reproducing kernel Hilbert space (RKHS). Typical examples of sparsity and smoothness terms are illustrated in Fig. 2.7.

Example 1 Given a measured response variable $y \in \mathbb{R}^m$ and a projection matrix $W \in \mathbb{R}^{m \times n}$, which represents the discrete version of Radon Transform, we need to reconstruct the latent input variable $x \in \mathbb{R}^n$.

$$y = f(x; w) = \sum_{i=1}^n w_i x_i + \eta = Wx + \eta, \quad \text{where } \eta \sim \mathcal{N}(0, \sigma),$$

where $\eta \in \mathbb{R}^n$ is i.i.d gaussian random variable with zero mean and σ standard deviation.

Solution: The latent variable x can be obtained by solving either for the least square error (LSE) or the maximum likelihood (ML),

$$x_{LSE} = \arg \min_x \frac{1}{2} \|y - Wx\|_2^2, \quad \text{or} \quad x_{ML} = \arg \max_x p(y|x),$$

where $p(y|x) \sim \mathcal{N}(Wx, I) = \exp^{-\frac{1}{2} \|y - Wx\|_2^2}$. It turns out solving the later formula for the negative log likelihood would give the same solution of least square error, and can be written in a closed form:

$$x_{LSE/ML} = (W^T W)^{-1} W^T y.$$

For an ill-posed problem, *i.e.* when W is under-determined, ill-conditioned, or $W^T W$ is a singular matrix, the solution is not unique. Fig. 2.8 shows the difference of error surface between a well- and ill-posed problems. It is obvious that the latter one has an infinite number of solutions.

As indicated previously, a sub-optimal solution can be obtained using regularization techniques that regularize the ill-posedness by incorporating a prior knowledge. For instance, a smoothness constraint

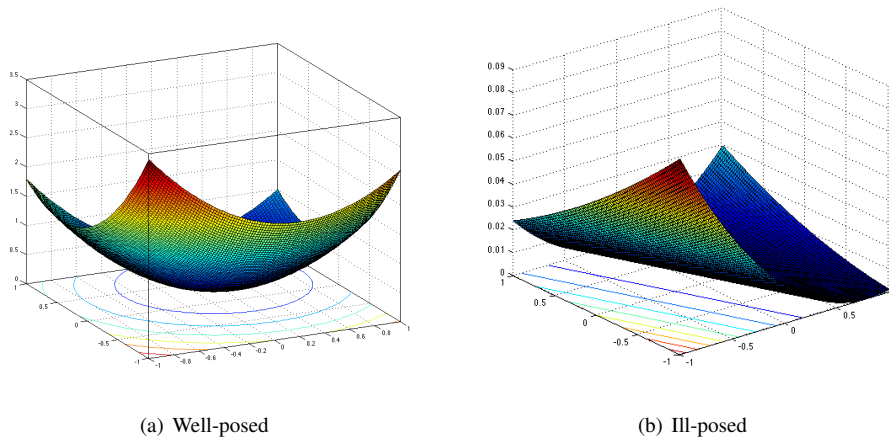


Figure 2.8. Error surface and contour of a well-posed vs. ill-posed.

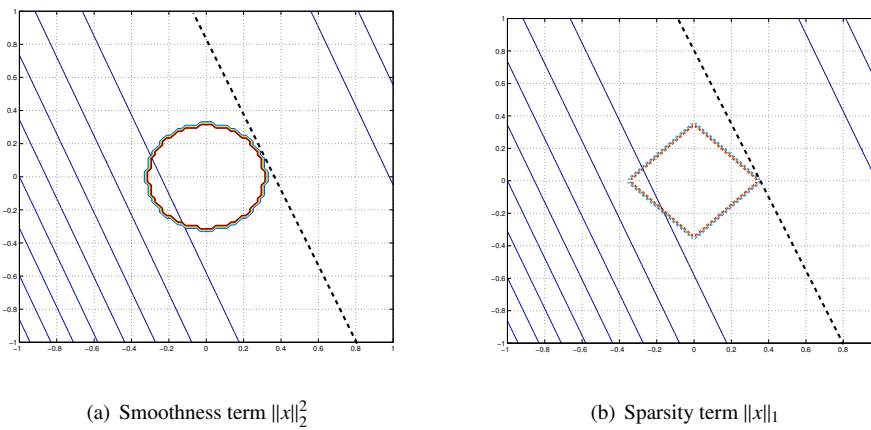


Figure 2.9. Sub-optimal solutions can be obtained using regularization techniques or soft constraints. Black dotted line shows the infinite number of solutions, where the colored shape represents the constraint. The intersection point is the sub-optimal solution.

on the latent variable can regularize the singular values and obtain a sub-optimal solution x_R analytically using Tikhonov regularization [97] or statistically using Maximum A Posterior (MAP) as follows:

$$x_R = \arg \min_x \frac{1}{2} \|y - Wx\|_2^2 + \frac{\lambda}{2} \|x\|_2^2, \quad \text{or} \quad x_{MAP} = \arg \max_x p(y|x) \cdot p(x),$$

where $p(x) \sim \mathcal{N}(0, I) = \exp^{-\frac{\lambda}{2} \|x\|_2^2}$ is the smoothness prior. Closed form solution can be written as

$$x_{R/MAP} = (W^T W + \lambda I)^{-1} W^T y.$$

2.2.3 Dictionary Learning

Unlike dimensionality reduction techniques such as Principle Component Analysis (PCA) where the data is projected into a lower dimensional using orthonormal basis of a complete dictionary, Dictionary

Learning is introduced in the context of compressed sensing and sparse coding, where the data is projected into a higher dimensional space using an overcomplete dictionary. Concretely, a signal $x \in \mathbb{R}^n$ is decomposed to a sparse linear combinations of basis elements $D \in \mathbb{R}^{n \times l}$, where $n \ll l$ as follows,

$$\arg \min_{D, \alpha} \frac{1}{2} \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \quad \text{s.t.} \quad \|d_i\|_2^2 = 1, \quad d_i \perp d_j, \quad (2.14)$$

where $\alpha \in \mathbb{R}^l$ is the sparse coefficient (code), λ is the regularization parameter, and $d_i \in \mathbb{R}^n$ is a column of the dictionary. The resulting sparse codes can be utilized as a feature vector for any un/supervised methods afterwards. Fig. 2.11 shows an embedding into a 3-dimensional space of raw features, deep learned features, and the sparse codes obtained from deep learned features.

Dictionary Learning has been intensively investigated for patch-based approaches in many applications for noise reduction [66], color image restoration [161], clustering and classification [199]. Further, it has been shown in the literature that sparse representation gives better compression performance over the complete dictionary [131], and more likely to be linearly separable [88]. Readers are referred to Aharon *et al.* [12] and Mairal *et al.* [159] for more details about the sparse coding and dictionary learning.

2.2.4 Deep Learning

In deep learning, particularly in Artificial Neural Networks (ANN), the predicted variable $\hat{\mathbf{y}} \in \mathbb{R}^m$ is modelled as a composite of nonlinear functions of input and hidden variables as

$$\hat{\mathbf{y}} = f(\mathbf{x}; \mathbf{w}) = f_{\mathbf{w}_h} \circ f_{\mathbf{w}_{h-1}} \circ \dots \circ f_{\mathbf{w}_0}(\mathbf{x}) \quad (2.15)$$

where each composite is a function of previous hidden variables (neurons) and associated parameters (weights) as

$$f_{\mathbf{w}_{h-1}}(\mathbf{h}) = \sigma(\mathbf{w}_{h-1}^T \mathbf{h} + b_{h-1}), \quad (2.16)$$

where $\sigma(\cdot)$ is as an activation function, $\mathbf{h} \in \mathbb{R}^u$ is the hidden variable (neurons), $\mathbf{w}_{h-1} \in \mathbb{R}^u$ are the associated learned parameters (weights), and b_{h-1} is the bias. Single layer of neural networks is referred to Perceptron (*cf.* Fig. 2.10).

Unlike ANN, Convolutional Neural Networks (CNN), which was first proposed by LeCun *et al.* [140] for handwritten character recognition, uses shared weights in the convolutional layers, followed by pooling layers, before it uses all neurons in the fully connected layers (similar to ANN). Each convolutional layer can be modelled as a convolution process of the hidden vector and the shared weights added to a certain bias as

$$f_{\mathbf{w}_{h-1}}(\mathbf{h}) = \sigma((\mathbf{w}_{h-1} * \mathbf{h}) + b_{h-1}), \quad (2.17)$$

where $\mathbf{h} \in \mathbb{R}^u$ is u -dimensional vector, $\mathbf{w}_{h-1} \in \mathbb{R}^v$ is v -dimensional vector, and $v \ll u$ (*cf.* Fig. 2.10). This architecture has way less parameters to estimate enabling us to handle 2-dimensional input data such as images. Nowadays, it is considered the state-of-the-art for object recognition in Computer Vision [133].

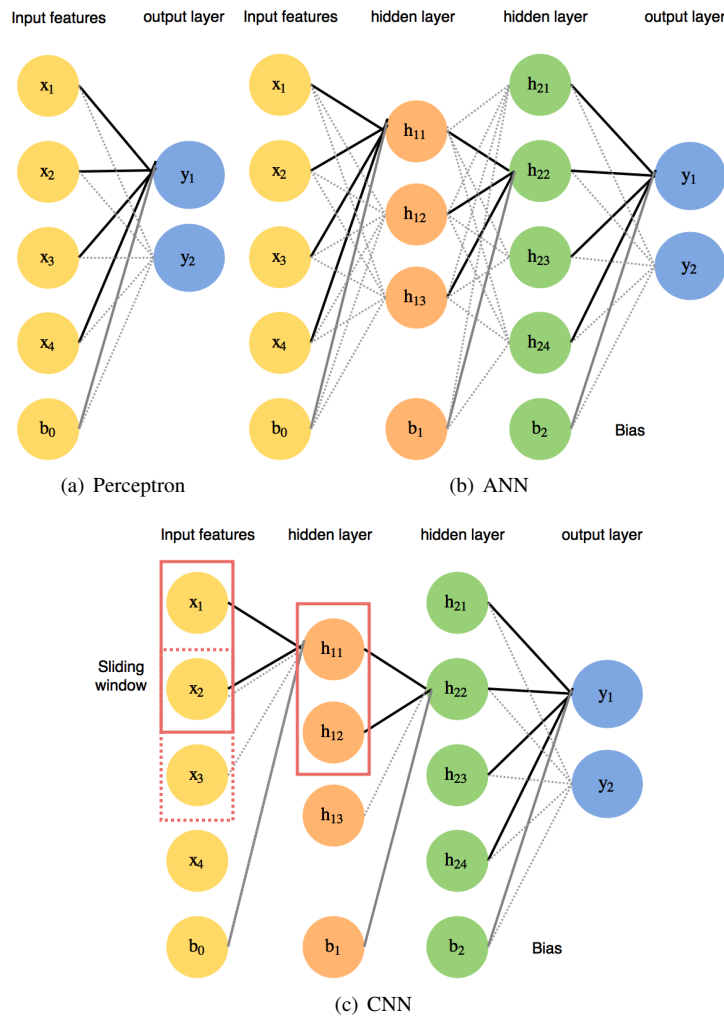


Figure 2.10. Different Feed-Forward Neural Networks Architectures

2.3 Evaluation

Over the last decades, researchers have investigated the evaluation metrics, in particular, the Receiver Operating Characteristics (ROC) curve, in radiology [186, 188, 291]. It has been shown that ROC and its Area Under the Curve (AUC) are fundamental tools to evaluate the model performance, particularly Computer Aided Diagnosis systems. In this section, a brief overview of the model selection is presented before reporting the common evaluation metrics of high interest for both classification and image quality in medical applications.

What are the criteria to select a ML model?

Free parameters available in the aforementioned machine learning algorithms govern the model complexity. For instance, the regularization parameter in linear regression controls the influence of the prior knowledge, and hence the model complexity, whereas for more complex models, such as convolutional neural networks, there are many free hyper-parameters, *i.e.* number of hidden layers and neurons, and learning rate, need to be carefully selected. The primary objective here is to find the

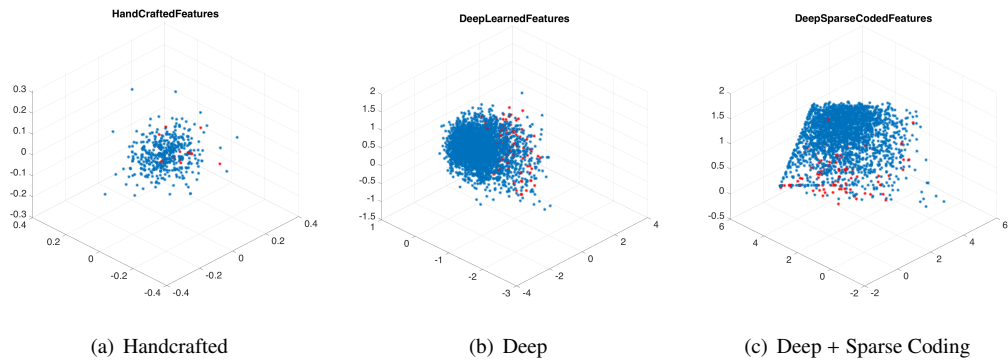


Figure 2.11. An embedding to 3-dimensional space of (a) Handcrafted features, (b) Deep Learned features, and (c) Deep Learned features encoded using Sparse Coding, for small patches of mitotic figures in histology images. The majority negative class in blue, and the minority positive class in red.

Table 2.2. Confusion (Decision) Matrix

	True Condition		
	Positive	Negative	
Prediction	TP	FP	T_+
Positive	FN	TN	T_-
Negative	D_+	D_-	
Total			

appropriate values of such hyper-parameters yielding the best model which can generalize well on new data. Therefore, such a model need to be trained several times for a different range of values to find the best one.

However, observing the performance on the training set is not a good indicator of the performance on a new unseen data due to the *overfitting* problem (cf. Sec. 2.2). Therefore, it is highly recommended to split the medical dataset into independent and preferably patient-wise; *training*, *validation*, and *testing* sets. For a limited amount of data, in particular for medical applications, the performance on the *validation* set is not entirely reliable. *Cross-validation* such as leave-p-out (LpOCV), leave-one-out (LOOCV), or k-fold cross validation, is one solution to obtain a reliable indicator of the model performance.

2.3.1 Classification Metrics

Before we explain the ROC and its importance to the medical diagnosis, the following metrics should be introduced first. Given a confusion matrix (cf. Table 2.2), the following metrics can be computed as follows:

Accuracy and Error Rate

Both are commonly used and reported to evaluate a classifier performance,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad ErrorRate = 1 - Accuracy, \quad (2.18)$$

where TP = True Positive, FP = False Positive, TN = True Negative, and FN = False Negative.

Precision

It is the percentage of cases that the classifier labeled as positive are actually positive, it is also known as positive predictive value (PPV). Thus, it is defined as

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{T_+}, \quad (2.19)$$

where the higher the precision, the more accurate the diagnosis.

Recall

It is the percentage of positive cases that the classifier did label as positive, it is also known as sensitivity or true positive rate (TPR). Thus, it is defined as

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{D_+}, \quad (2.20)$$

where the higher the recall, the more accurate the diagnosis.

Specificity

It is the percentage of negative cases that the classifier did label as negative, it is also referred as true negative rate (TNR). Thus, it is defined as

$$Specificity = \frac{TN}{TN + FP} = \frac{TN}{D_-}, \quad (2.21)$$

where the higher the specificity, the more accurate the diagnosis.

F-Measure

It is the harmonic mean of precision and recall, and provides more insight into the functionality of the classifier. Therefore, it is more effective than accuracy, in particular for class imbalance.

$$F_\beta = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \quad (2.22)$$

where β controls the balance between precision and recall (*cf.* Fig. 2.12). Typical value of β is 1 or 2. The higher the F -score, the better the overall performance.

Receiver Operating Characteristics (ROC)

As indicated earlier, ROC is an effective and fundamental method in the evaluation of model performance. It is defined as a plot of model sensitivity or TPR as the y-coordinate against its 1-specificity or False Positive Rate (FPR) as the x-coordinate, where both sensitivity and specificity are computed at every possible threshold. One of the most popular metrics is the area under the ROC curve (AUC), where the higher the AUC, the better the overall performance. Four ROC curves are illustrated in Fig. 2.12, whereas curves C and D show the random guessing (AUC= 0.5) and the perfect (AUC \approx 1.0) performance, respectively, both curves A and B exhibits the same AUC. However, to determine the better model among A and B, the clinician may select the optimal operating point (black circles in Fig. 2.12) based on the clinical application. For imbalanced classes, Precision-Recall curve is highly recommended [112].

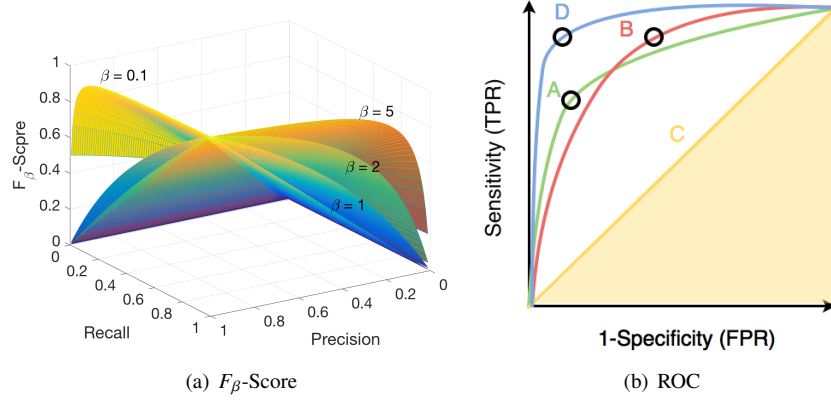


Figure 2.12. (a) Mesh surface of F_β -Score at different values of β . (b) ROC curves of different models.

Example 2 Given the following confusion matrix, $TP = 50$, $FP = 10$, $TN = 9840$, and $FN = 50$, compute the evaluation metrics?

Answer: Accuracy = 99.4%, Precision = 83.33%, Recall = 50%, F_1 -score = 62.50%. As anticipated, Accuracy is not a reliable metric in the presence of imbalance data.

2.3.2 Image Quality Metrics

Few metrics are reported here. However, readers are referred to this review [189] for more details.

Mean Square Error (MSE)

It is a pixel-wise measure and commonly used in reconstruction and noise reduction. MSE is given as

$$MSE(x, y) = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2, \quad (2.23)$$

where x is the predicted or regressed pixel. The lower the MSE, the better the quality.

Peak Signal-to-Noise Ratio (PSNR)

It is a measure of the peak error between the predicted image and original one, PSNR is given as

$$PSNR(x, y) = 10 \log_{10} \left(\frac{y_{max}^2}{MSE} \right), \quad (2.24)$$

where y_{max} is the maximum value of the image. The higher the PSNR, the better the quality.

Structural Similarity (SSIM)

It is a measure of perceptual image quality [266] and given as

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (2.25)$$

where μ is the mean, σ is the standard deviation, and C are constants. The higher the SSIM, the better the quality.

From Nano- to Milli-meter Image Resolution

” *The good physician treats the disease; the great physician treats the patient who has the disease!*

— William Osler (1849–1919)

3.1 Cryo-Electron Tomography

In contrast to Computed Tomography (CT), a medical device used to scan the interior of a patient, where the source-detector is arranged in such a way that it is rotated around the patient along a single axis (*cf.* Fig 3.1), Electron Tomography (ET) uses Electron source to image the interior of tiny objects that commonly used for biological imaging [151] and material science [93]. The source of illumination, the angular or the spatial resolution regarding the wavelength, is defined by the Rayleigh criterion. The shorter wavelength, the greater resolution is obtained. Hence, the Electron beam is used as the imaging source to visualize the details of the molecular structures (around $4 - 10\text{\AA}$). These Electrons, accelerated by the voltage imposed on anode and cathode of the electron gun, travel in a vacuum tube and then pass through the specimen. The scattered and unscattered electrons are then collected to form a 2D projection image on the film surface. By rotating the sample around its axis, different projection images can be obtained. Unlike, the Electron Microscopy (EM) which is utilized for Single Particle Analysis (SPA) such as macromolecules, Electron Tomography (ET) enables the understanding and interpretation of three-dimensional complex cellular structures [123].

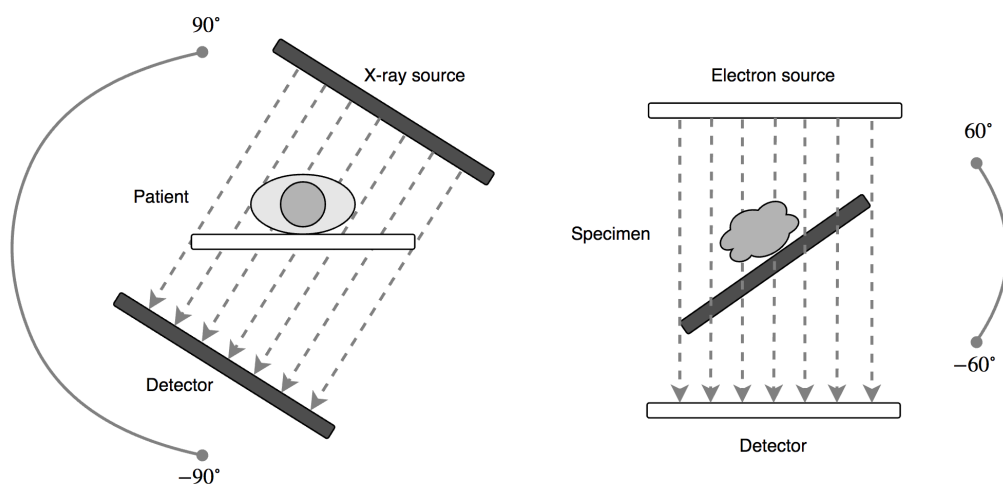


Figure 3.1. Difference between Computed Tomography (left) and Electron Tomography (right). Stationary parts in white while the rotating ones in black. Limited angle view is available in ET.

Different techniques such as [16] and [272] have been developed towards native-state imaging in biological contexts such as Cryo-Electron Microscopy of vitreous sections (CEMOVIS), which has become an important technique for structural molecular biology at cryogenic temperatures (*cf.* Fig. 3.2). Over the last decade, cryo-Electron Tomography (CET) has drawn the attention of researchers. It is considered the most powerful imaging technique to address fundamental questions on biological structures at both cellular and molecular levels [170] (*cf.* Fig. 3.2). It also bridges the gap between low-resolution imaging techniques, *e.g.* light microscopy, high-resolution techniques, *e.g.* single particle electron microscopy, X-ray crystallography, and nuclear magnetic resonance (NMR), to get better understanding and more insights into the mechanism of protein structures and viruses [33], hence, aids drug design development [19].

3.1.1 Problem Definition and Motivation

Cryo-ET merges the principles of transmission electron microscopy (TEM) and the principle of tomographic imaging by acquiring several two-dimensional projection images of biological structures at limited tilt range and close-to-native condition. These two-dimensional projection images are then reconstructed into a three-dimensional image (called tomogram), after passing through a pipeline of alignment and restoration procedure as shown in Fig. 3.3. Ideally, these aligned 2D projections can be used to analytically reconstruct the underlying 3D structure using Weighted Back-Projection (WBP) techniques, which assumes for each tilt angle, the projection image represents the mass density encountered by the image rays. Accordingly, it redistributes the known specimen mass that is present over the back-projection rays, so that the specimen mass is projected back into the reconstruction volume. This is done overall tilt angles, and the reconstructed volumes are 'strongly blurred' where the mass is frequently present. This process is assisted by the use of weighting (filtering) techniques to mitigate the blurring effects. Iterative reconstruction techniques along with compressive sensing techniques have shown better and interesting results for Image reconstruction, particularly for limited angle view [41]. Once the latent structure of the 3D tomographic object is reconstructed, a post-processing step, *i.e.* noise reduction, tomograms averaging, and segmentation, becomes necessary to interpret the underlying structure. For more in-depth description of CET and the associated image processing pipeline, see [81]. While it carries much hope and potential use-cases to the community, it is still hampered by mechanical and technical limitations calling for novel and advanced techniques on both hardware and software solutions. Next, few limitations and challenges present in CET such as the limited angle, radiation dose, and EM transfer function, are briefly introduced.

Limited Angle

Due to the physical limitation of ET, the data collected at angular range restricted between $\pm 60^\circ/70^\circ$ around the tilting axis causing what so-called missing wedge effect. This missing information which can be easily observed in the Fourier domain of the 3D volume as a missing cone (thanks to Central Section Theorem), negatively affects the reconstruction yielding elongated and blurred objects [81]. Further, it makes the 3D Inverse Fourier Transform challenging since the radial frequency is not an isotropic anymore.

Radiation Damage

An excessive amount of radiation dose may cause a specimen damage. Therefore the maximum tolerable dose of the electron beam is divided by the number of 2D projections. This adds a trade-off between the number of projections and the electron dose per projection, generating an extremely

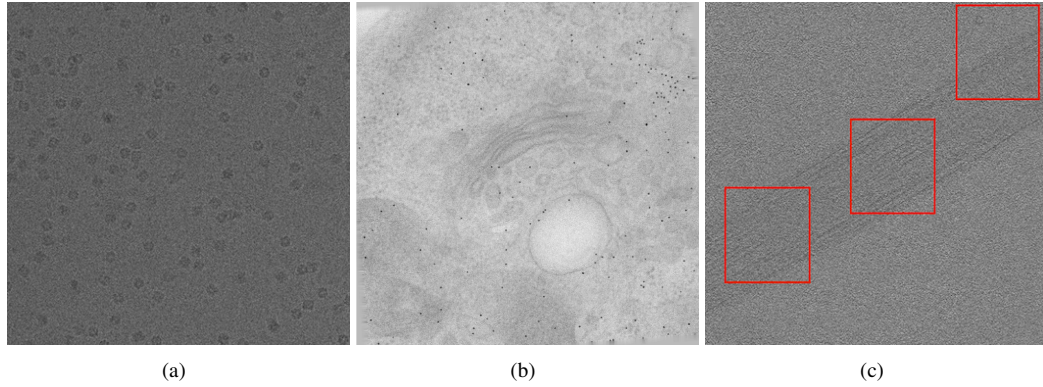


Figure 3.2. Electron Microscopic Images of (a) GroEL for Single Particle Analysis and (b) Hella cells for Cellular structures. A single 2D projection of a sperm cell scanned by Cryo-Electron Tomography(c) . Black dots appeared in (b) are gold particles used for feature-based alignment methods. Red boxes show the region of interests. Images (a) and (b) courtesy of Electron Microscopy Group and BirkBeck College, respectively.

low signal to noise ratio ($SNR < 0.1$). While the resulting noise distribution can be modeled as a Poisson noise added to the 2D projections, it is different for the 3D tomogram as it is highly dependant on the reconstruction algorithms [81]. For simplicity, it is commonly defined as a Gaussian distribution [101].

EM Contrast Transfer Function (CTF)

Another problem related to the imaging device is the CTF, which is defined as an oscillating function of the spatial frequency and profoundly affected by the applied defocus and other parameters of the imaging system. Since CTF function crosses the zero line behaving as a sinusoidal function, the contrast of many frequencies is flipped (negative lobes), and others are lost (zero crossing) affecting the contrast and further the resolution of the reconstructed tomograms. To mitigate this problem, few methods are proposed for the CTF correction [279] (*cf.* Fig. 3.3).

Given the previous challenges, advanced image processing methods are highly desirable, particularly for 3D reconstruction and noise reduction, to reduce the noise, and compensate the missing wedge effect while at the same time preserving the structure of interest.

3.1.2 Related Work

It has been shown that iterative tomographic reconstruction algorithms perform better than the back-projection techniques yielding an improvement in the resolution and contrast. Standard techniques such as Simultaneous Iterative Reconstruction Technique (SIRT) or Simultaneous Algebraic Reconstruction Technique (SART) are introduced, where the solution can be obtained using least squares approach using

$$\mathbf{x}_{LS} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{b} - A\mathbf{x}\|_2^2, \quad (3.1)$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\mathbf{x} \in \mathbb{R}^n$ is the three-dimensional reconstructed tomogram, $\mathbf{b} \in \mathbb{R}^m$ represents the measured projection data, and $A \in \mathbb{R}^{m \times n}$ represents the weighting matrix, where a_{ji} is the weight associated with which each voxel in the image vector $\mathbf{x} \in \mathbb{R}^n$ contributes to the j -th projection measurement.

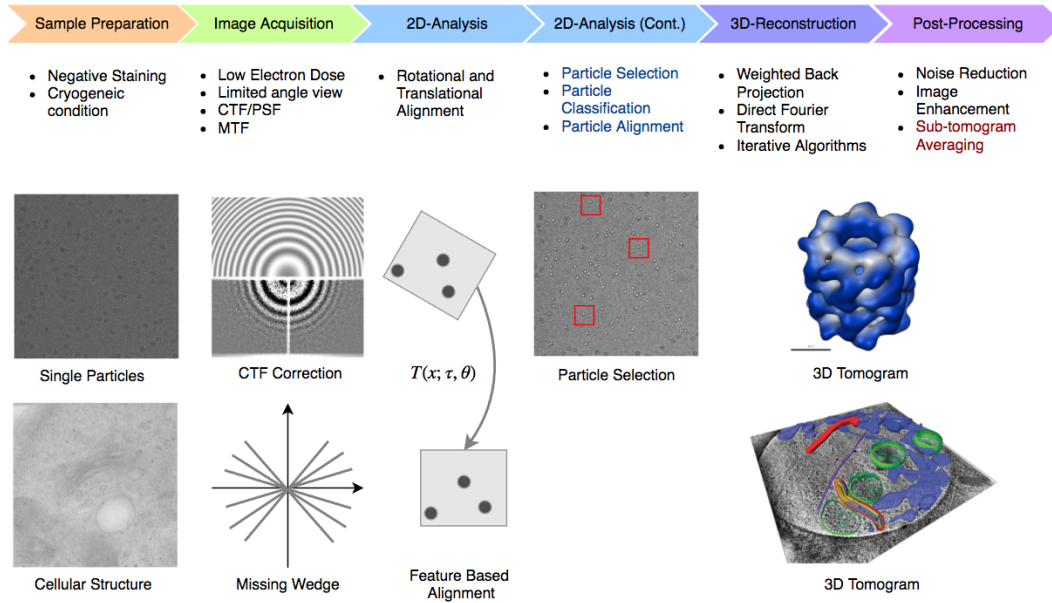


Figure 3.3. Image Processing Pipeline for Single Particle Analysis (top) and cellular structures (bottom). Specific tasks for Single Particle Analysis (in blue) and for Cellular Structures (in red). 3D Tomogram of Cellular structure images courtesy of National Institute of Medical Research.

Table 3.1. Different SIRT techniques

Method	T	M
Landweber	I	I
Cimminos	I	$\frac{1}{m} \text{diag}(\frac{1}{\ a^i\ _2^2})$
Component Averaging (CAV)	I	$\frac{1}{m} \text{diag}(\frac{1}{\ a^i\ _s^2})$
SART	$\text{diag}(\frac{1}{\ a^i\ _1})$	$\text{diag}(\frac{1}{\ a^i\ _1})$

This least squares problem can be solved iteratively following a weighted gradient descent approach,

$$\mathbf{x}^{k+1} = \mathbf{x}^k + s_k \mathbf{g}_k, \quad k = 0, 1, \dots, \quad (3.2)$$

with a starting value \mathbf{x}^0 , step size s_k and a weighted gradient $\mathbf{g}_k = TA^T M(\mathbf{b} - A\mathbf{x}^k)$, where T and M are diagonal matrices represented by the sum of the columns or rows of the system matrix A . A different definition of these matrices leads to different Simultaneous Iterative Reconstruction Technique (SIRT) methods [205] (cf. Table 3.1). This applies also to the recently developed techniques I-SIRT [100], M-SART [246] or W-SIRT [267]. However, due to the strong measurement noise and the limited number of projections, a least squares approach might fail and lead to noise amplification. As pointed out in a previous example (cf. Ex. 1), regularization-based techniques that incorporate prior knowledge are introduced in the context of tomographic reconstruction in particular for medical imaging, *i.e.* Computed Tomography, showing promising results [222]. In very recent work, Compressed Sensing is investigated as prior knowledge to improve the reconstruction of Cryo-Electron Tomographic data [101].

Table 3.2. Related work on Tomographic Reconstruction and Noise Reduction in CET.

Paper	Task	Methodology
Liu <i>et al.</i> , 2009 [253]	Tomographic Reconstruction	Modified SART
Guo <i>et al.</i> , 2012 [102]	Tomographic Reconstruction	Improved SIRT
Wolf <i>et al.</i> , 2014 [275]	Tomographic Reconstruction	Weighted SIRT
Fernandez <i>et al.</i> , 2003 [75]	Noise Reduction	Improved Anisotropic Nonlinear Diffusion
Darbon <i>et al.</i> , 2008 [50]	Noise Reduction	Non-Local Means Filter
Narasimha <i>et al.</i> , 2008 [178]	Noise Reduction	Evaluation of different denoising methods
Fernandez <i>et al.</i> , 2009 [74]	Noise Reduction	Beltrami flow
Fleet <i>et al.</i> , 2014 [77]	Noise Reduction	Rolling Guidance Filter

3.1.3 Contribution: Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction (CMMI MICCAI 2014)

By introducing Huber term as a smooth regularization term to the energy function, we have constrained the solution to a feasible bounded set. Also, such a smooth term is preferred over non-smooth regularization terms, *i.e.* Isotropic total variation, that might pose a problem during the optimization procedures. Moreover, our proposed regularized energy function can be easily optimized using projected gradient methods (see Appendix A). We have validated our proposed reconstruction method against commonly used methods in Electron Tomography, including the filtered back projection and the algebraic reconstruction techniques, on a vitrified freeze-substituted section of HeLa cells. Results have been evaluated using Fourier Shell Correlation, and Line Profile, showing an out-performing result of our proposed method.

3.1.4 Contribution: Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data (BMVC 2015)

By introducing Graph Laplacians, with a full control of scale-space and global consistency, as a regularization term to the energy function, we were able to reduce the noise in CET data without over smoothing fine-scale structures. Overlapped batches (voxels) are collected from images (volumes) and treated as nodes in the graph, where the weights are computed based on the distance between these patches on a multi-scale pyramid, *i.e.* where the noise manifests itself at coarse levels, and the hidden structures are revealed (see Appendix B). It turned out that the core component in the closed form solution is the graph spectral filter (GSF), which controls the frequency decay and the degree of smoothness present in the Laplacian Graphs. We have validated our proposed algorithm on Computer Vision as well as CET data. Our proposed algorithm significantly outperforms the state-of-the-art methods regarding noise removal and structure preservation.



Figure 3.4. Histology sample preparation. Image courtesy of AMIDA 2013 MICCAI Grand Challenge.

3.2 Breast Cancer Histology Images

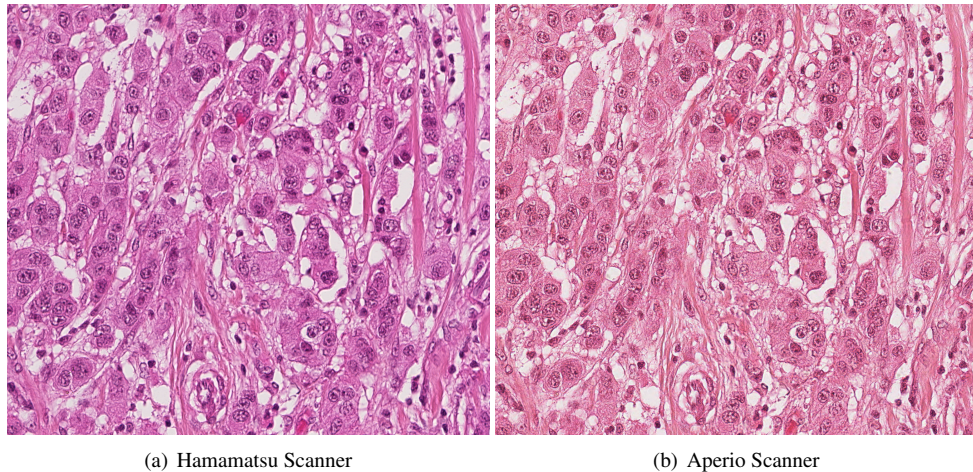
One of the leading women death's causes in the developed and developing countries is breast cancer, approximately 1.7 million cases and 521,900 deaths in 2012 [231]. In the USA, based on the American Cancer Society's estimate¹, about 300,000 new cases will be diagnosed of invasive or non-invasive breast cancer in 2017. Many risk factors raise the incidence of breast cancer in the developing world including increases in smoking, excess body weight, and physical inactivity [237]. Although some control measures and risk reduction strategies are applied, the majority of breast cancer is diagnosed in later stages. Therefore, early breast cancer detection is essential to prevent any progression and to increase the survival rate.

Regular check-ups, using mammography or ultrasound imaging, are performed to detect and diagnose the breast cancer in early stages. Once the exam indicates a possibility of an abnormal lesion growth, a breast tissue biopsy is undertaken to confirm the diagnosis. Collected tissue during the biopsy is sliced into small cuts and thin sections (*cf.* Fig. 3.4). Then, these sections are put onto glass slides, before they are stained with hematoxylin and eosin (H&E).

3.2.1 Problem Definition and Motivation

To identify and detect mitotic figures, the specimen is investigated under the light microscopy, nowadays with the digital slide scanners, where pathologists screen the whole slide image (WSI). Once the most active region is identified, the pathologists start counting the mitotic figures available in an area of 2 sqm, sub-divided into 10 patches, so-called high power field (HPF) slides (*cf.* Fig. 3.6). A typical case takes around 5 – 10 min. for a pathologist to perform mitosis counting. Based on the number of detected mitotic figures, the cancer is graded accordingly [208]. Readers are referred to the modified Bloom–Richardson–Elston grading (BRE) [68] system for more details about breast cancer classification. Next, we discuss briefly few present challenges in breast cancer histology imaging.

¹<https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>



(a) Hamamatsu Scanner

(b) Aperio Scanner

Figure 3.5. Color variations of the same stained tissue among different scanners. Image courtesy of MITOS-ATYPIA-14 ICPR Challenge.

H&E Staining

In histopathology, the hematoxylin and eosin stain is the most common standard protocol used. That is because of its simplicity and ability to enhance the appearance and increase the contrast between different tissue structures. It is simply achieved by binding these stains, which have different dyes, to specific proteins. For instance, hematoxylin, in combination with mordant, binds strongly to the nuclei showing a dark purple color, while eosin dye binds to the cytoplasm, stroma, and other structures, showing a pink color. Nevertheless, undesirable variations in RGB colors can vary widely due to many factors, *i.e.* raw materials, manufacturing techniques of stain vendors, and different digital slide scanners (*cf.* Fig.3.5). Technically, this variation between scanners, so-called domain shift, creates complexity for any automatic machine learning models trained on a particular stain appearance. Recently, stain normalization methods [153, 10, 11] are developed to mitigate this problem.

Inter- and Intra-Observer Variability

Another challenge presents in histology imaging, in particular Breast Cancer, is the high variability between Inter- and Intra-observers in the diagnosis of atypical ductal hyperplasia (ALH), Lobular (LCIS), and Ductal (DCIS) carcinoma in situ (CIS) of the breast. Therefore, Dice coefficient and Cohen’s Kappa coefficient are calculated to analyze the agreement between pathologists. In Gomes *et al.* study [89], the inter-observer variability between general pathologists and an expert in breast pathology was investigated, a weak correlation was observed for the diagnosis of different types of carcinoma (average Kappa = 0.47). In a recent study, Elmore *et al.* [67] noted that the overall agreement between individual pathologists and the expert consensus was about 75.3%. In both studies, higher agreement of invasive cancer, and lower levels of understanding of CIS were observed. Overall, these studies highlighted the need for automated models to reduce this variability.

Given the challenges mentioned above, an automatic and reproducible method for detection of mitotic figures, in breast cancer histology images, has a great potential to assist pathologists, reduce the total amount of time and efforts, and mitigate the inter- and intra-observer variability.

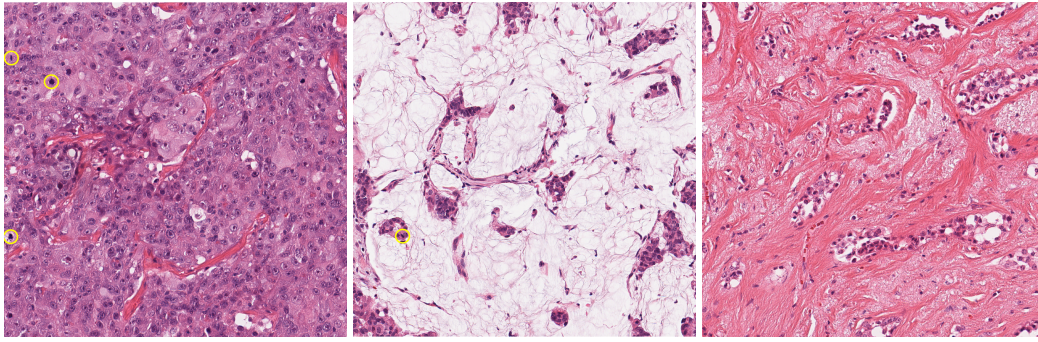


Figure 3.6. Detected mitotic figures, in yellow circles, on different HPF slides. Image courtesy of AMIDA 2013 MICCAI Grand Challenge.

3.2.2 Related Work

In recent years, with the increased availability of WSI scanners, many grand challenges, *i.e.* AMIDA13² and CAMELYON16³, called for (semi-)automated solutions to detect mitotic figures [246] and metastasis [25] in breast cancer, respectively. Classical ML algorithms based on engineered and handcrafted features are proposed to detect mitotic figures, however, among the participants, Ciresan *et al.* [44] proposed a deep max-pooling convolutional neural network. Their proposed patch-based CNN model set the state-of-the-art, and won both the ICPR 2012 competition and the AMIDA 13 challenge (*cf.* Fig. 3.7). Readers are referred to Veta *et al.* [246] for more details.

One interesting paragraph in Veta *et al.* [246] paper quoted here,

“After a visual inspection of the detection results, it was observed that many of the false positives produced by the top performing methods closely resemble mitotic figures. Indeed, owing to the difficulty of the task it is possible that some mitotic figures were missed during the ground truth annotation, but were then detected by the automatic methods.”

This interesting finding sparked the idea of investigating the collaboration of both Human Intelligence, via Crowdsourcing, and Artificial Intelligence, by using Deep Learning frameworks.

3.2.3 Contribution: AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images (IEEE TMI 2016)

By integrating Crowdsourcing into Deep Learning framework, we were able to i) mitigate the negative influence of noisy crowd votes (spammers), ii) generate so-called crowd-truth annotations using a robust aggregation layer, and iii) improve the performance of CNN models.

An initial multi-scale CNN model is proposed to detect the mitotic figures at different scale-space, then the probabilistic scores are geometrically averaged to produce more accurate detections. The proposed

²<http://amida13.isi.uu.nl/>

³<http://camelyon16.grand-challenge.org/>

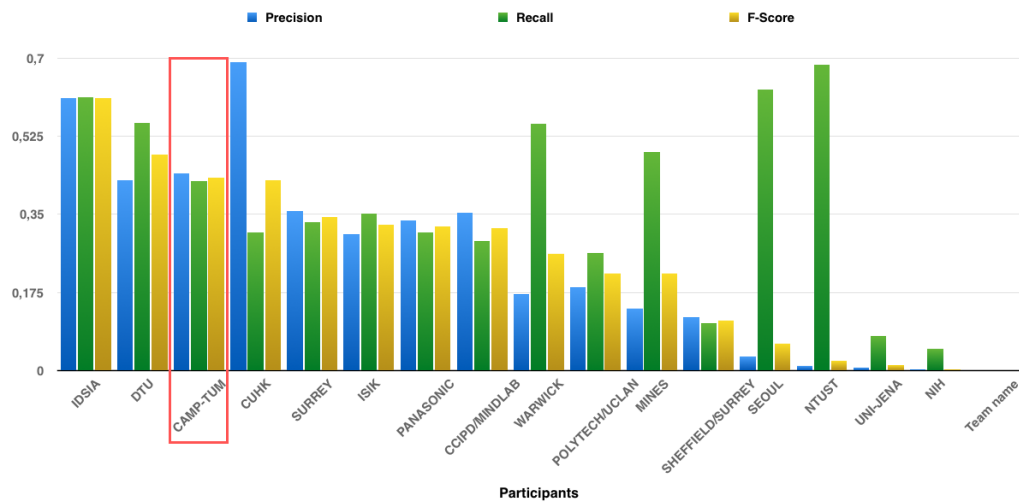


Figure 3.7. Ranking of all participants in AMIDA13 Challenge. Our approach rank is 3 out of 16. The full team name can be found in the challenge web page.

model has been validated on a publicly available database, namely AMIDA13 challenge, showing a comparable performance to the state-of-the-art method (*cf.* Fig. 3.7).

To validate the primary objectives of this work, a subset of the training set is treated as a validation set, where the detected mitotic figures, obtained from the pre-trained CNN model, are sent to the crowdsourcing platform. Tutorials along with some examples were provided to the participants. Besides, quality control was performed to allow only serious participants. Collected crowd votes were fed back to the network through a novel robust aggregation layer (see Appendix C). Interestingly, the model that fine-tuned from aggregated crowd votes showing an outstanding performance. This novel approach has been published in a *Special Issue in Deep Learning* in the IEEE Transactions on Medical Imaging.

3.2.4 Contribution: Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research (LABELS/DLMIA MICCAI 2016)

By introducing a novel concept of an image to game-object translation in Biomedical Imaging, we have been able to represent detected mitotic figure images to star-shaped objects that can be embedded easily to any readily available game canvas. The novel concept targets non-expert users for crowdsourcing tasks via gamification providing an incentive for persistent engagement of the players. In our work, we have noted an interesting and promising result of our playsourcing concept compared to the other conventional crowdsourcing platforms (see Appendix D).

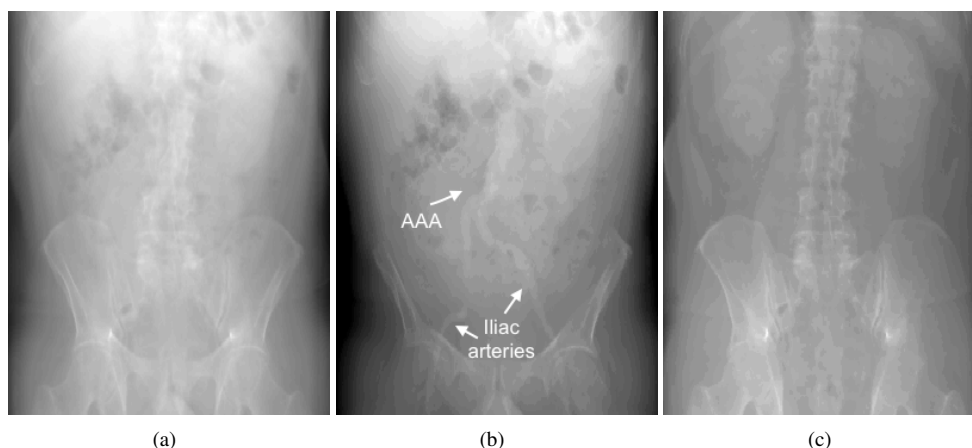


Figure 3.8. Simulated X-ray images for (a) abdominal: $\vec{\mu}_T$, (b) AAA sub-volume: $\vec{\mu}_A$, and (c) spine sub-volume: $\vec{\mu}_S$. Region of interests (white arrows) are quite visible after removing rigid structure, *i.e.* bone.

3.3 Depth Perception in Interventional X-ray Imaging

Since its discovery by Roentgen in 1895, X-ray radiology is still the primary imaging modality for diagnostic and interventional radiology. Due to its projective nature, a single pixel in an X-ray image may partially contribute to soft tissue as well as hard tissue encoding the depth information within the convolution of attenuation and scale. Hence, the accumulation of X-ray attenuation displayed on the X-ray image leads to anatomy obscuration. In abdominal X-ray imaging, observing and distinguishing the anterior and posterior organs; spine, pelvis, and the vasculature, *i.e.* Abdominal Aorta Aneurysm (AAA), requires skilled clinicians (*cf.* Fig. 3.8).

In Minimally Invasive Surgery (MIS), mobile C-arm X-ray imaging device has proven to be an essential part of the surgical workflow and has been widely used, to assist clinicians and surgeons, in many medical procedures [115]. The acquired intra-operative fluoroscopy or angiography images are used together with the pre-operative 3D information to better understand the underlying anatomy. To this end, many methods and techniques are introduced to improve the visual perception of obscured anatomical structures, *i.e.* contrast-enhanced X-ray [173], dual-energy X-ray [169], Dual-energy contrast-enhanced X-ray [143], and phase contrast X-ray [31, 289], just to name a few. Neither of the methods above provides a distinction between the anatomy at different depth and yet distinguishing the layers of the anterior and posterior anatomy depends on the surgeon's experience and judgment. Next, we briefly present the impact of perceptual errors on health care, before we discuss few loosely related works that depend mainly on multi-modal 3D/2D registration.

Cognitive and Perceptual Errors in Radiology

In last decades, errors in radiology gained a lot of attention, from the community, calling for proposed solutions to prevent medical errors and build safer healthcare systems. In 1999, the US Institute of Medicine [60] reported that around 98,000 patients die every year in hospitals due to diagnostic errors that could have been prevented. Very similar data has been published by Hayward *et al.* [111]. Whereas Pinto *et al.* [195] proposed quality improvement projects in both knowledge and systems, Graber *et al.* [94] argued that being able to measure the incidence of diagnostic errors is essential to initiate such quality improvement projects. In a very recent review, Adrian Brady [30] highlighted both human-

and system-derived errors in details, reporting both cognitive and perceptual errors, which account for 20 – 40% and 60 – 80% in total, respectively. His conclusion was that no single strategy can eliminate error in radiology. Overall, these studies suggested the use of CAD systems to reduce the perceptual errors and mitigate the inter- and intra-variability between radiologists.

Multi-modal 3D/2D Registration

To improve the perceptual problem in 2D X-ray radiographs, prior knowledge from pre-operative patient CT data is used to restore depth data. However, as pointed out by Terry Peters [194], accurate multi-modal registration between 3D pre-operative CT scan and 2D X-ray image is required. Wieczorek *et al.* [274] have demonstrated that estimating the relative depth of the X-ray image relative to the CT data allows to modulate X-ray intensities (cutting off at a certain depth) and to introduce additional depth cues into the X-ray image. In [258], an interactive virtual mirror is integrated into the 2D/3D fusion of X-ray and CT data to facilitate the localization of an aneurysm. The virtual mirror generates desired perspectives of the 3D data and displays 2D and 3D data on a single screen. To further improve the depth perception of X-ray acquisitions, colored depth maps are computed using ray casting on 3D data, and are later integrated into the interventional X-ray while preserving the original gray-scale level of the X-ray [256, 264]. These methods rely on accurate 2D/3D registration of data which by itself is a challenging task given the highly dynamic environment, *i.e.* motion due to respiration, and organ deformation. Last but not least, access to patient CT data is not available for the majority of cases.

Given the challenges above, correct depth recovery of the underlying anatomy, from a single view X-ray image, would have a great potential to assist radiologists and surgeons in both diagnostic and interventional procedures.

3.3.1 Related Work

Whereas the task of depth recovery from single-view images has been investigated thoroughly within the computer vision community, few attempts have been studied in Image-Guided Intervention (IGI), in particular for robotic-assisted MIS [158, 185, 227]. In Computer Vision, learning-based approaches, where both monocular images (features) together with the corresponding depth images are used for training an ML model, to recover the depth information for a given monocular image, yielded successful applications for 3D reconstruction [216], and scene understanding [263]. In robotic-assisted MIS, Stoyanov *et al.* proposed interesting methods for depth recovery and scene reconstruction [227, 229], where they made use of the available information of camera position, in addition to the known 3D geometry.

3.3.2 Contribution: Single-view X-ray depth recovery: toward a novel concept for image-guided interventions (IJCARS 2016)

By training a patient specific model from a pre-operative 3D CT scan, we have been able to i) roughly estimate the corresponding pose of the C-arm device, and ii) recover the corresponding depth information of a given fluoroscopic X-ray image during the intervention.

Employing the recently proposed gradient-based rendering scheme [264], we have been able to generate ground-truth depth images for any X-ray image simulated at different possible C-arm configurations (*cf.* Fig. 3.9). Our proposed depth model employs both depth images and the corresponding X-ray

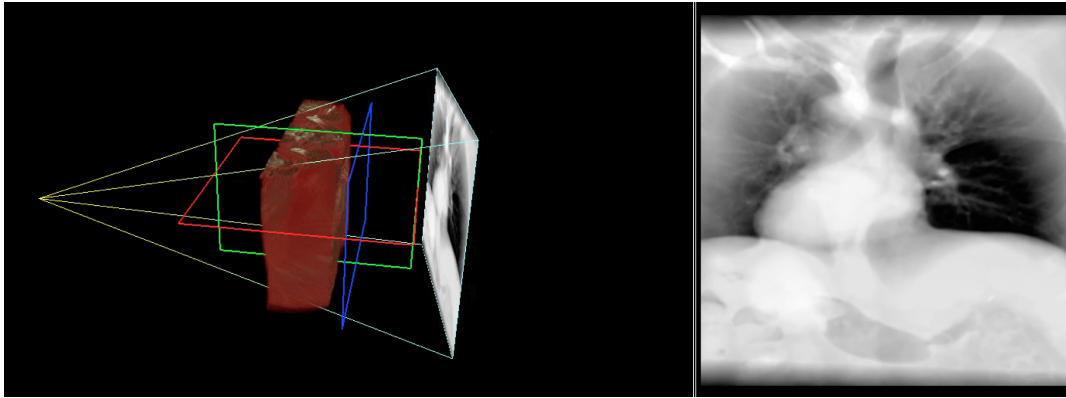


Figure 3.9. Generated Digitally Reconstructed Radiographs (DRR) from the soft-tissue sub-volume.

images to train a label consistent dictionary. Using the trained dictionary, a rough depth estimation is obtained for any given query image. To improve the recovered recovery further, an atlas prior together with spatial information is utilized (see Appendix E).

3.3.3 Contribution: X-ray In-Depth Decomposition: Revealing The Latent Structures (MICCAI 2017)

By incorporating the Beer-Lambert Law as a constraint, we have been able to decompose a given X-ray image into d in-depth layers separating Rigid Structures, *i.e.* spine, and ribcage, into distinct layers, leaving all Deformable Organs in one layer.

Our proposed deep learning model is trained on a couple of thousands of pairs of generated DRRs of the whole CT volume, along with the corresponding DRRs of individual, independent, non-overlapped in-depth sub-volumes (*cf.* Fig. 3.9). The highly ill-posed regression problem is constrained by a fundamental principle of X-ray imaging, *i.e.* Beer-Lambert Law, to be able to ensemble the given X-ray image.

We have validated our proposed model on two clinical databases showing a promising result. Further, we have demonstrated an impressive result on a clinical use case for Tuberculosis detection. Two CAD models have been trained individually using i) the conventional chest X-ray image, and ii) the deformable organs, *i.e.* Lung, and Vasculatures, showing an encouraging result for the later one (see Appendix F).

Conclusion and Outlook

Throughout this dissertation, a prior knowledge, *i.e.* domain-specific knowledge, has been investigated for many biomedical applications illustrating a positive influence on the performance. In this context, a detailed overview of the major challenges in machine learning for medical application was provided before focusing on the contributions. Now, we want to summarize our contributions addressing the posed research questions:

- **To what extent can Artificial Intelligence (AI) and Human Intelligence (HI) collaborate for robust Biomedical Image Annotation?**

In this context, we have presented a novel concept for learning from crowds, *i.e.* training an AI from HI, that can robustly aggregate the *noisy* annotations, collected via Crowdsourcing platform, during the training process. It has been confirmed that HI can work together with AI positively influencing the model performance provided that i) the initial AI model is trained using ground-truth data, ii) quality control is applied.

We have also introduced a novel concept for Playsourcing in the biomedical context, where we gamify the crowdsourcing task for medical applications. Our framework is designed to transform images into a visual salient game object that can be embedded easily into a readily available game canvas. It has been shown that Playsourcing performs better than crowdsourcing ones, reducing the cost, time, and false positives.

One more research question needs further investigation:

- *Can we eliminate the need for the initial model using ground-truth labels and rather train a model from scratch using the noisy annotations?*

We have noticed in our experiments on Breast Cancer Histology Images that it is nearly impossible due to the high inter- and intra-variability between pathologists and even worse for nonexpert participants. That might be possible for simpler tasks, *i.e.* organ localization, and segmentation, however, this need to be further investigated. One suggestion is to train an initial model in unsupervised fashion, *i.e.* AutoEncoders, and do transfer learning afterward.

- **How can prior knowledge be defined and incorporated into Machine Learning (ML) algorithms?**

It has been shown that the definition of a prior knowledge mainly depends on a thorough understanding of the clinical problem and close interaction with clinicians and practitioners. This prior knowledge has been modeled as a soft constraint, *i.e.* regularization term, in the corresponding energy function.

- **What are the challenges present in real scenarios, how this would affect the evaluation, and what are the possible solutions?**

Throughout the dissertation, many biomedical applications have been presented highlighting the significant challenges that negatively influence the performance. For instance, one of the major challenges available in Breast Cancer Histology Images is the highly imbalanced classes and the domain shift. To mitigate that in the preprocessing phase, data augmentation and stain normalization have been employed, respectively. In later works, not discussed in this dissertation, a cost-sensitive energy function [7], and an auxiliary manifold embedding [8], have been introduced to mitigate the class imbalance and domain shift problems, respectively. Interestingly, it has been noted that training a deep learning model on a huge amount of DRRs can be transferred to real X-ray images for Depth Perception application. This observation has been confirmed in our recent research on pose estimation for X-ray images [9].

In our contributions to both challenges, the assessment of mitosis detection algorithms (AMIDA13) and the cancer metastasis detection in lymph node (CAMELYON16), we have introduced a multi-scale deep learning approach. Results are evaluated using robust evaluation metrics showing an exciting performance compared to other participants.

List of Figures

1.1	Energy sources and different imaging modalities.	4
1.2	Google trends on Big Data and Artificial Intelligence; Pattern Recognition, Machine Learning, and Deep Learning.	4
2.1	Venn Diagram of Data Science: Drew Conway [218] (left), Modified one (right). . . .	8
2.2	MLBiD framework. Adapted from Fig.1 in [288].	8
2.3	Class Imbalance in (a) classical scenarios, and (b) complex ones with the presence of intra- and inter-class imbalance. Example on synthetic methods, and its drawback in (c) and (d).	9
2.4	Big data or Small data: On top is the learning healthcare system proposed by Sacristan <i>et al.</i> [212], while the bottom figure, adapted from [72], shows the long tail of unpublished and dark datasets, which can be a potential treasure for big data.	11
2.5	Bias-Variance tradeoff: dotted lines refers to the desired performance, where the model complexity refers to the number of iterations in this context.	12
2.6	Two cultures: a) Data Modeling, b) Algorithmic Modeling.	13
2.7	The geometry of different norms in 3-dimensional space: Sparsity prior in (a-b), and Smoothness prior in (c-d).	15
2.8	Error surface and contour of a well-posed vs. ill-posed.	16
2.9	Sub-optimal solutions can be obtained using regularization techniques or soft constraints. Black dotted line shows the infinite number of solutions, where the colored shape represents the constraint. The intersection point is the sub-optimal solution.	16
2.10	Different Feed-Forward Neural Networks Architectures	18
2.11	An embedding to 3-dimensional space of (a) Handcrafted features, (b) Deep Learned features, and (c) Deep Learned features encoded using Sparse Coding, for small patches of mitotic figures in histology images. The majority negative class in blue, and the minority positive class in red.	19
2.12	(a) Mesh surface of F_β -Score at different values of β . (b) ROC curves of different models.	21
3.1	Difference between Computed Tomography (left) and Electron Tomography (right). Stationary parts in white while the rotating ones in black. Limited angle view is available in ET.	23
3.2	Electron Microscopic Images of (a) GroEL for Single Particle Analysis and (b) Hella cells for Cellular structures. A single 2D projection of a sperm cell scanned by Cyro-Electron Tomography(c) . Black dots appeared in (b) are gold particles used for feature-based alignment methods. Red boxes show the region of interests. Images (a) and (b) courtesy of Electron Microscopy Group and BirkBeck College, respectively.	25

3.3	Image Processing Pipeline for Single Particle Analysis (top) and cellular structures (bottom). Specific tasks for Single Particle Analysis (in blue) and for Cellular Structures (in red). 3D Tomogram of Cellular structure images courtesy of National Institute of Medical Research.	26
3.4	Histology sample preparation. Image courtesy of AMIDA 2013 MICCAI Grand Challenge.	28
3.5	Color variations of the same stained tissue among different scanners. Image courtesy of MITOS-ATYPIA-14 ICPR Challenge.	29
3.6	Detected mitotic figures, in yellow circles, on different HPF slides. Image courtesy of AMIDA 2013 MICCAI Grand Challenge.	30
3.7	Ranking of all participants in AMIDA13 Challenge. Our approach rank is 3 out of 16. The full team name can be found in the challenge web page.	31
3.8	Simulated X-ray images for (a) abdominal: $\vec{\mu}_T$, (b) AAA sub-volume: $\vec{\mu}_A$, and (c) spine sub-volume: $\vec{\mu}_S$. Region of interests (white arrows) are quite visible after removing rigid structure, <i>i.e.</i> bone.	32
3.9	Generated Digitally Reconstructed Radiographs (DRR) from the soft-tissue sub-volume.	34
B.1	MG²F Framework: A noisy image slice from the 3D reconstructed tomogram is fed to the algorithm, where the graph is built on a selected scale space image (<i>i.e.</i> coarse grid) acting as a guidance for the regularized graph spectral filter.	64
B.2	Spectral filters responses: (a) linear (<i>i.e.</i> Bilateral), (b) regularised graph, and (c) designed one, against the parameters λ (spectral frequency) and α (regularization parameter), (d) shows the line profile ($\alpha = 1$) for different filters.	66
B.3	Photographic Image: Results of different algorithms on Lena image(128X128, SNR=7) along with a tabulated comparison to the proposed MG ² F filter.	68
B.4	Sensitivity analysis: PSNR contour against k-NN and Patch size for different image sizes (a) 64 ² and (b) 128 ² . In (c) PSNR of different denoising algorithms (NLM, NAD, RGF and Ours respectively) for 150 slices (SNR=0.1) from 15 simulated tomograms. (d) FSC curve for different denoised 3D tomograms.	69
B.5	2D CET data: Filtering results on the tomogram along with the corresponding CNR of b) NLM (0.1979), c) NAD (0.2570), d) RGF (0.3146), e) Proposed MG ² F (0.3150), where the arrows point to the fine structures on the membrane and the ellipse contains the inner core of HIV virus.	69
B.6	3D CET Data: A comparison between different 3D filtering methods to our proposed MG ² F method on real unstained HIV-1 data (EMDB-ID: 1155).	70
C.1	AggNet Framework: (1) The multi-scale CNN model is trained from gold-standard annotations. (2) Then for any incoming unlabelled image, (3) the <i>AggNet</i> will produce a response map which is thresholded at selected optimal operating point. (4) These few resulting positive candidates are outsourced to crowds. (5) <i>AggNet</i> collects back the crowd votes and jointly aggregates the ground truth and refine the CNN model.	74
C.2	AggNet architecture: The same CNN architecture is used for different scales, where p_i, μ_i, y_i^j represents the classifier output, the aggregated label, and the crowdvotes respectively.	78
C.3	Instructions and guidelines	80
C.4	First row shows results of one single image using multi-scale CNN, Green: the true positives, Orange: the false positives. Second row shows the corresponding final detection map (FDM) before thresholding. Best viewed in color.	81
C.5	Evaluation Metrics: Precision, Recall, and F_1 -score of Patients 9, 11 and 12.	82

C.6	(a) The aggregated labels of the crowdsourcing set are evaluated using the F_1 -score metric. (b) The loss function barely changes at 3-8 epochs before starting to overfit and the gap between the validation and training curves becomes significant. The shaded area depicts the change of τ	84
C.7	ROC curves of the (a) aggregated labels using MV, GLAD and the proposed AG-NoQ, (b) the augmented models AM-GT, AM-MV, AM-GLAD and <i>AggNet</i> as well.	85
C.8	Participants Analysis: accuracy and spammer scores of 100 participants. Arrows in green show some participants achieve high accuracy scores in the qualitative test, however, they are spammer. Arrows in red show very few participants who have good accuracy score as well as spammer score. Note that spammer score "0" means the participant is spammer.	87

List of Tables

2.1	Publicly available challenges and the rank obtained by the author's team.	12
2.2	Confusion (Decision) Matrix	19
3.1	Different SIRT techniques	26
3.2	Related work on Tomographic Reconstruction and Noise Reduction in CET.	27
C.1	DATASETS SPECIFICATIONS	79
C.2	F_1 -SCORES	82
C.3	AGGREGATED LABELS	83
C.4	AUGMENTED MODELS	83
C.5	AGGREGATION AND DETECTION RESULTS	86
C.6	USE CASE RESULTS	86
C.7	USE CASE AUGMENTED MODELS	87

Bibliography

- [12] M. Aharon, M. Elad, and A. Bruckstein. “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation”. In: *IEEE Trans Signal Proc* 54.11 (2006), pp. 4311–4322 (cit. on p. 17).
- [13] A. Aichert, M. Wiczorek, J. Wang, et al. “The Colored X-rays”. In: *Proc. MICCAI-Workshop on Augmented Environments for Computer-Assisted Interventions*. Ed. by C. A. Linte, E. C. Chen, M.-O. Berger, J. T. Moore, and D. R. H. III. Vol. 7815. LNCS. Springer, 2012, pp. 45–54.
- [14] S. Aksoy and R. M. Haralick. “Feature normalization and likelihood-based similarity measures for image retrieval”. In: *Pattern recognition letters* 22.5 (2001), pp. 563–582 (cit. on p. 8).
- [15] T. Aksoy, G. B. Unal, S. Demirci, N. Navab, and M. Degertekin. “Template-based CTA to X-ray Angio Rigid Registration of Coronary Arteries in Frequency Domain with Automatic X-Ray Segmentation”. In: *Med Phys* 40.10 (2013).
- [16] A. Al-Amoudi, J.-J. Chang, A. Leforestier, et al. “Cryo-electron microscopy of vitreous sections”. In: *The EMBO journal* 23.18 (2004), pp. 3583–3588 (cit. on p. 24).
- [17] J. Al-Khalili. “The first ‘true scientist’”. In: *Accessed on January 30* (2009), p. 2013.
- [1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. “AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1313–1321.
- [2] S. Albarqouni, T. Lasser, W. Alkhaldi, A. Al-Amoudi, and N. Navab. “Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction”. In: *Computational Methods for Molecular Imaging*. 2015, pp. 43–51.
- [3] S. Albarqouni, M. Baust, S. Conjeti, A. Al-Amoudi, and N. Navab. “Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data.” In: *British Machine Vision Conference (BMVC)*. 2015, pp. 17–1.
- [4] S. Albarqouni, S. Matl, M. Baust, N. Navab, and S. Demirci. “Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research.” In: *LABELS/DLMIA@ MICCAI*. 2016, pp. 269–277.
- [5] S. Albarqouni, U. Konrad, L. Wang, N. Navab, and S. Demirci. “Single-view X-ray depth recovery: toward a novel concept for image-guided interventions”. In: *International journal of computer assisted radiology and surgery* (2016), pp. 1–8.
- [6] S. Albarqouni, J. Fotouhi, and N. Navab. “X-ray In-Depth Decomposition: Can Deep Learning Reveal The Latent Structures?” In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017*. Vol. Lecture Notes in Computer Science Volume 10435. 2017, pp. 401–409.
- [18] Alhazen, A.-H. I. Al-Haytham, and A. Sabra. *Kitab al-Manazir*. National Council for Culture, Arts and Letters, 1983 (cit. on p. 3).
- [19] A. M. Andrianov. “Computational anti-AIDS drug design based on the analysis of the specific interactions between immunophilins and the HIV-1 gp120 V3 loop. Application to the FK506-binding protein”. In: *Journal of Biomolecular Structure and Dynamics* 26.1 (2008), pp. 49–56 (cit. on p. 24).

- [20] L. Armijo. “Minimization of functions having Lipschitz continuous first partial derivatives”. In: *Pacific Journal of mathematics* 16.1 (1966), pp. 1–3.
- [21] L. Aroyo and C. Welty. “The Three Sides of CrowdTruth”. In: *Human Computation* 1.1 (2014), pp. 31–44.
- [22] L. Aroyo and C. Welty. “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation”. In: *AI Magazine* 36.1 (2015) (cit. on p. 73).
- [23] X.-C. Bai, G. McMullan, and S. H. Scheres. “How cryo-EM is revolutionizing structural biology”. In: *Trends in biochemical sciences* 40.1 (2015), pp. 49–57.
- [24] L. Barrington, D. Turnbull, and G. Lanckriet. “Game-powered machine learning”. In: *Proceedings of the National Academy of Sciences* 109.17 (2012), pp. 6411–6416.
- [7] C. Baur, S. Albarqouni, S. Demirci, N. Navab, and P. Fallavollita. “CathNets: Detection and Single-View Depth Prediction of Catheter Electrodes”. In: *International Conference on Medical Imaging and Virtual Reality*. Springer. 2016, pp. 38–49 (cit. on p. 36).
- [8] C. Baur, S. Albarqouni, and N. Navab. “Semi-Supervised Learning for Fully Convolutional Networks”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2017*. Vol. Lecture Notes in Computer Science Volume 10435. 2017, pp. 281–289 (cit. on p. 36).
- [25] B. E. Bejnordi, M. Veta, M. Hermsen, et al. “Analysis: Deep learning outperforms pathologists in detecting lymph node metastases of breast cancer: the CAMELYON16 challenge”. In: *Submitted to Journal of the American Medical Association (JAMA)* (2017) (cit. on p. 30).
- [26] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. “Robust Optimization for Deep Regression”. In: *Proc. IEEE Int. Conf. Computer Vision (ICCV)*. IEEE. 2015 (cit. on p. 73).
- [27] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006 (cit. on p. 7).
- [28] L. Bottou. “How big data changes statistical machine learning”. In: *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE. 2015, pp. 1–1 (cit. on p. 10).
- [29] L. Bottou. “Large-scale machine learning with stochastic gradient descent”. In: *Proc. Computational Statistics*. 2010, pp. 177–186 (cit. on p. 77).
- [30] A. P. Brady. “Error and discrepancy in radiology: inevitable or avoidable?” In: *Insights into Imaging* 8.1 (2017), pp. 171–182 (cit. on p. 32).
- [31] A. Bravin, P. Coan, and P. Suorti. “X-ray phase-contrast imaging: from pre-clinical applications towards clinics”. In: *Physics in medicine and biology* 58.1 (2012), R1 (cit. on p. 32).
- [32] L. Breiman et al. “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. In: *Statistical Science* 16.3 (2001), pp. 199–231 (cit. on p. 13).
- [33] J. A. Briggs, K. Grünewald, B. Glass, F. Förster, H.-G. Kräusslich, and S. D. Fuller. “The mechanism of HIV-1 core assembly: insights from three-dimensional reconstructions of authentic virions”. In: *Structure* 14.1 (2006), pp. 15–20 (cit. on p. 24).
- [9] M. Bui, S. Albarqouni, M. Schrapp, N. Navab, and S. Ilic. “X-Ray PoseNet: 6 DoF Pose Estimation for Mobile X-Ray Devices”. In: *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE. 2017, pp. 1036–1044 (cit. on p. 36).
- [34] P. H. Calamai and J. J. Moré. “Projected gradient methods for linearly constrained problems”. In: *Mathematical programming* 39.1 (1987), pp. 93–116.
- [35] S. Candemir, S. Jaeger, K. Palaniappan, et al. “Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration”. In: *IEEE transactions on medical imaging* 33.2 (2014), pp. 577–590.
- [36] G. Carneiro, J. C. Nascimento, and A. Freitas. “The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods”. In: *IEEE Trans. Image Process.* 21.3 (2012), pp. 968–982 (cit. on p. 73).

- [37] G. Carneiro, T. Peng, C. Bayer, and N. Navab. “Weakly-Supervised Structured Output Learning With Flexible and Latent Graphs Using High-Order Loss Functions”. In: *Proc. IEEE Int. Conf. Computer Vision*. 2015, pp. 648–656.
- [38] L. A. Celi, A. Ippolito, R. A. Montgomery, C. Moses, and D. J. Stone. “Crowdsourcing knowledge discovery and innovations in medicine”. In: *J. Med. Internet Res.* 16.9 (2014) (cit. on p. 72).
- [39] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. “Return of the devil in the details: Delving deep into convolutional nets”. In: *British Machine Vision Conference (BMVC)*. 2014.
- [40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357 (cit. on p. 9).
- [41] G.-H. Chen, J. Tang, and S. Leng. “Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets”. In: *Medical physics* 35.2 (2008), pp. 660–663 (cit. on p. 24).
- [42] Y. Chen and F. Förster. “Iterative reconstruction of cryo-electron tomograms using nonuniform fast Fourier transforms”. In: *Journal of structural biology* 185.3 (2014), pp. 309–316.
- [43] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. “Mitosis detection in breast cancer histology images with deep neural networks”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. 2013, pp. 411–418 (cit. on p. 73).
- [44] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. “Mitosis detection in breast cancer histology images with deep neural networks”. In: (2013) (cit. on p. 30).
- [45] U. Clarenz, M. Droske, and M. Rumpf. “Towards fast non-rigid registration”. In: *Inverse Problems, Image Analysis, and Medical Imaging*. Ed. by M. Z. Nashed and O. Scherzer. AMS, 2002.
- [46] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.
- [47] S. Cooper, F. Khatib, A. Treuille, et al. “Predicting protein structures with a multiplayer online game”. In: *Nature* 446 (2010), pp. 756–760.
- [48] M. Daly, J. Siewerdsen, Y. Cho, D. Jaffray, and J. Irish. “Geometric calibration of a mobile C-arm for intraoperative cone-beam CT”. In: *Med Phys* 35.5 (2008), pp. 2124–2136.
- [49] A. Danielyan, V. Katkovnik, and K. Egiazarian. “BM3D frames and variational image deblurring”. In: *Image Processing, IEEE Transactions on* 21.4 (2012), pp. 1715–1728 (cit. on p. 67).
- [50] J. Darbon, A. Cunha, T. Chan, S. Osher, and G. Jensen. “Fast nonlocal filtering applied to electron cryomicroscopy”. In: *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*. 2008, pp. 1331–1334 (cit. on pp. 27, 62, 66, 68).
- [51] P. R. DeLucia, R. D. Mather, J. A. Griswold, and S. Mitra. “Toward the improvement of image-guided interventions for minimally invasive surgery: three factors that affect performance”. In: *Hum Factors* 48.1 (2006), pp. 23–38.
- [52] S. Demirci, O. Kutter, F. Manstad-Hulaas, R. Bauernschmitt, and N. Navab. “Advanced 2D-3D registration for Endovascular Aortic Interventions: Addressing Dissimilarity in Images”. In: *Proc. SPIE Medical Imaging*. 2008, 69182S–69190S.
- [53] S. Demirci, M. Baust, O. Kutter, F. Manstad-Hulaas, H.-H. Eckstein, and N. Navab. “Disocclusion-based 2D-3D Registration for Angiographic Interventions”. In: *Computers in Biology and Medicine* 43.4 (2013), pp. 312–322.
- [54] S. Demirci, A. Bigdelou, L. Wang, et al. “3D stent recovery from one x-ray projection”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011, pp. 178–185.
- [55] T. M. Deserno, H. Handels, H.-P. Meinzer, et al. “On Feature Tracking in X-Ray Images”. In: *Proc. Bildverarbeitung für die Medizin (BVM)*. Springer Berlin Heidelberg, 2014, pp. 132–137.

- [56] S. Didas, B. Burgeth, A. Imiya, and J. Weickert. “Regularity and Scale-Space Properties of Fractional High Order Linear Filtering.” In: *Scale-Space*. Ed. by R. Kimmel, N. A. Sochen, and J. Weickert. Vol. 3459. Lecture Notes in Computer Science. Springer, 2005, pp. 13–25 (cit. on p. 65).
- [57] C. DIEBOLDER, A. J. Koster, R. I. Koning, et al. “Pushing the resolution limits in cryo electron tomography of biological structures”. In: *Journal of microscopy* 248.1 (2012), pp. 1–5.
- [58] K. Doi. “Computer-aided diagnosis in medical imaging: historical review, current status and future potential”. In: *Computerized medical imaging and graphics* 31.4 (2007), pp. 198–211 (cit. on p. 4).
- [59] P. Domingos. “A few useful things to know about machine learning”. In: *Communications of the ACM* 55.10 (2012), pp. 78–87.
- [60] M. S. Donaldson, J. M. Corrigan, L. T. Kohn, et al. *To err is human: building a safer health system*. Vol. 6. National Academies Press, 2000 (cit. on p. 32).
- [61] A. Dumitrache, L. Aroyo, C. Welty, R.-J. Sips, and A. Levas. “Dr. Detective: combining gamification techniques and crowdsourcing to create a gold standard in medical text”. In: *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web-Volume 1030*. CEUR-WS. org, 2013, pp. 16–31.
- [62] M. Dumoux, D. K. Clare, H. R. Saibil, and R. D. Hayward. “Chlamydiae assemble a pathogen synapse to hijack the host endoplasmic reticulum”. In: *Traffic* 13.12 (2012), pp. 1612–1627.
- [63] D. Eigen and R. Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture”. In: *arXiv preprint arXiv:1411.4734* (2014).
- [64] D. Eigen, C. Puhrsch, and R. Fergus. “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network”. In: *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*. 2014, pp. 2366–2374.
- [65] D. Eigen, C. Puhrsch, and R. Fergus. “Depth map prediction from a single image using a multi-scale deep network”. In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2014 (cit. on p. 73).
- [66] M. Elad and M. Aharon. “Image denoising via sparse and redundant representations over learned dictionaries”. In: *IEEE Transactions on Image processing* 15.12 (2006), pp. 3736–3745 (cit. on p. 17).
- [67] J. G. Elmore, G. M. Longton, P. A. Carney, et al. “Diagnostic concordance among pathologists interpreting breast biopsy specimens”. In: *Jama* 313.11 (2015), pp. 1122–1132 (cit. on p. 29).
- [68] C. W. Elston and I. O. Ellis. “Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up”. In: *Histopathology* 19.5 (1991), pp. 403–410 (cit. on p. 28).
- [69] E. Estellés-Arolas and F. González-Ladrón-De-Guevara. “Towards an Integrated Crowdsourcing Definition”. In: *J. Inf. Science* 38.2 (2012), pp. 189–200 (cit. on p. 72).
- [70] A. Esteva, B. Kuprel, R. A. Novoa, et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639 (2017), pp. 115–118 (cit. on p. 4).
- [71] P. Fallavollita, A. Winkler, S. Habert, et al. “Desired-View Controlled Positioning of Angiographic C-arms”. In: *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2014, pp. 659–66.
- [72] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone. “Big data from small data: data-sharing in the ‘long tail’ of neuroscience”. In: *Nature neuroscience* 17.11 (2014), pp. 1442–1447 (cit. on pp. 10, 11).
- [73] J. Fernandez, S. Li, and R. Crowther. “CTF determination and correction in electron cryotomography”. In: *Ultramicroscopy* 106.7 (2006), pp. 587–596.
- [74] J.-J. Fernandez. “TOMOBFLOW: feature-preserving noise filtering for electron tomography”. In: *BMC bioinformatics* 10.1 (2009), p. 178 (cit. on pp. 27, 68).

- [75] J.-J. Fernández and S. Li. “An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms”. In: *Journal of structural biology* 144.1 (2003), pp. 152–161 (cit. on p. 27).
- [76] M. A. Figueiredo, R. D. Nowak, and S. J. Wright. “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems”. In: *Selected Topics in Signal Processing, IEEE Journal of* 1.4 (2007), pp. 586–597.
- [77] D. Fleet, T. Pajdla, B. Schiele, et al. “Rolling Guidance Filter”. In: *Computer Vision – ECCV 2014*. Vol. 8691. Lecture Notes in Computer Science. Springer, 2014, pp. 815–830 (cit. on pp. 27, 62, 63, 66, 68).
- [78] A. Foncubierta Rodríguez and H. Müller. “Ground truth generation in medical imaging: a crowdsourcing-based iterative approach”. In: *Proc. ACM Multim. Work. Crowdsourcing for Multimedia*. 2012, pp. 9–14 (cit. on p. 72).
- [79] D. Forsyth, P. Torr, A. Zisserman, A. Vedaldi, and S. Soatto. “Quick Shift and Kernel Methods for Mode Seeking”. In: *Proc. European Conference on Computer Vision (ECCV)*. Springer Berlin Heidelberg, 2008, pp. 705–718.
- [80] J. Frank. *Electron Tomography Methods for Three-Dimensional Visualization of Structures in the Cell*. Springer, 2006 (cit. on pp. 62, 66, 67).
- [81] J. Frank. *Electron tomography: methods for three-dimensional visualization of structures in the cell*. Springer Science & Business Media, 2008 (cit. on pp. 24, 25).
- [82] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini. “The influence of contour on similarity perception of star glyphs”. In: *Visualization and Computer Graphics, IEEE Transactions on* 20.12 (2014), pp. 2251–2260.
- [83] B. C. Fung, K. Wang, and S. Y. Philip. “Anonymizing classification data for privacy preservation”. In: *IEEE transactions on knowledge and data engineering* 19.5 (2007) (cit. on p. 10).
- [84] S. Gabarda and G. Cristóbal. “Blind image quality assessment through anisotropy”. In: *J. Opt. Soc. Am. A* 24.12 (2007), B42–B51.
- [85] A. Gadde, S. Narang, and A. Ortega. “Bilateral filter: Graph spectral interpretation and extensions”. In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*. 2013, pp. 1222–1226 (cit. on pp. 62, 63, 65).
- [86] A. G. Gallagher, P. P. Kearney, K. J. McGlade, L. B. Lonn, and G. C. O’Sullivan. “Avoidable Factors Can Compromise Image-Guided Interventions”. In: *Medscape General Surgery* (2012).
- [87] R. Girshick, J. Donahue, T. Darrell, and J. Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 2014 (cit. on p. 73).
- [88] X. Glorot, A. Bordes, and Y. Bengio. “Deep Sparse Rectifier Neural Networks.” In: *Aistats*. Vol. 15. 106. 2011, p. 275 (cit. on p. 17).
- [89] D. S. Gomes, S. S. Porto, D. Balabram, and H. Gobbi. “Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast”. In: *Diagnostic pathology* 9.1 (2014), p. 121 (cit. on p. 29).
- [90] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye. “A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems”. In: *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*. Vol. 28. 2. NIH Public Access. 2013, p. 37.
- [91] B. M. Good and A. I. Su. “Crowdsourcing for bioinformatics”. In: *Bioinformatics* (2013), btt333.
- [92] B. M. Good, S. Loguercio, O. L. Griffith, M. Nanis, C. Wu, and A. I. Su. “The Cure: design and evaluation of a crowdsourcing game for gene selection for breast cancer survival prediction”. In: *JMIR Serious Games* 2.2 (2014).

- [93] B Goris, T. Roelandts, K. Batenburg, H. H. Mezerji, and S Bals. “Advanced reconstruction algorithms for electron tomography: from comparison to combination”. In: *Ultramicroscopy* 127 (2013), pp. 40–47 (cit. on p. 23).
- [94] M. L. Graber. “The incidence of diagnostic error in medicine”. In: *BMJ Qual Saf* (2013), bmjqs–2012 (cit. on p. 32).
- [95] L. J. Grady and J. R. Polimeni. *Discrete Calculus, Applied Analysis on Graphs for Computational Science*. Springer, 2010 (cit. on p. 64).
- [96] R. Grimm, S. Bauer, J. Sukkau, J. Hornegger, and G. Greiner. “Markerless estimation of patient orientation, posture and pose using range and pressure imaging”. English. In: *Int J Comp Assist Radiol Surg* 7.6 (2012), pp. 921–929.
- [97] C. Groetsch. “The theory of Tikhonov Regularization for Fredholm Equations”. In: *104p, Boston Pitman Publication* (1984) (cit. on p. 16).
- [98] M. Groher, D. Zikic, and N. Navab. “Deformable 2D-3D registration of vascular structures in a one view scenario”. In: *IEEE Trans Med Imag* 28.6 (2009), pp. 847–860.
- [99] M. Groher, T. F. Jakobs, N. Padoy, and N. Navab. “Planning and Intraoperative Visualization of Liver Catheterizations: New CTA Protocol and 2D-3D Registration Method”. In: *Acad Radiol* 14.11 (2007), pp. 1325–1340.
- [100] C. G. Gross. “Ibn-al-Haytham on eye and brain, vision and perception”. In: *Bull. Islamic Med* 1 (1981), pp. 309–312.
- [101] M. D. Guay, W. Czaja, M. A. Aronova, and R. D. Leapman. “Compressed sensing electron tomography for determining biological structure”. In: *Scientific reports* 6 (2016) (cit. on pp. 25, 26).
- [102] W. Guo and H. Chen. “Improving SIRT algorithm for computerized tomographic image reconstruction”. In: *Recent Advances in Computer Science and Information Engineering*. Springer, 2012, pp. 523–528 (cit. on p. 27).
- [103] D. Gurari, D. Theriault, M. Sameki, et al. “How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-Experts, and Algorithms”. In: *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*. IEEE. 2015, pp. 1169–1176 (cit. on p. 72).
- [104] B. ter Haar Romeny, L. Florack, J. Koenderink, M. Viergever, and J. Weickert. “A review of nonlinear diffusion filtering”. In: *Scale-Space Theory in Computer Vision*. Vol. 1252. Springer Berlin Heidelberg, 1997, pp. 1–28 (cit. on p. 68).
- [105] J. Hadamard. “Sur les problèmes aux dérivées partielles et leur signification physique”. In: *Princeton university bulletin* (1902), pp. 49–52.
- [106] J. Hamari, J. Koivisto, and H. Sarsa. “Does gamification work?—a literature review of empirical studies on gamification”. In: *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. IEEE. 2014, pp. 3025–3034.
- [107] D. Hammond, Y. Gur, and C. Johnson. “Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel”. In: *IEEE Global Conf. on Sig. and Inf. Proc.* 2013, pp. 419–422 (cit. on p. 65).
- [108] C. M. Hampton, J. D. Strauss, Z. Ke, et al. “Correlated fluorescence microscopy and cryo-electron tomography of virus-infected or transfected mammalian cells”. In: *nature protocols* 12.1 (2017), pp. 150–167 (cit. on p. 10).
- [109] P. C. Hansen and M. Saxild-Hansen. “AIR tools—a MATLAB package of algebraic iterative reconstruction methods”. In: *Journal of Computational and Applied Mathematics* 236.8 (2012), pp. 2167–2178.
- [110] Z. Harmany, D. Thompson, R. Willett, and R. F. Marcia. “Gradient projection for linearly constrained convex optimization in sparse signal recovery”. In: *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE. 2010, pp. 3361–3364.

- [111] R. A. Hayward. “Counting deaths due to medical errors”. In: *JAMA* 288.19 (2002), pp. 2404–2404 (cit. on p. 32).
- [112] H. He and E. A. Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284 (cit. on pp. 9, 20).
- [113] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE on CVPR*. 2016, pp. 770–778.
- [114] M. Hein and M. Maier. “Manifold denoising”. In: *Advanced in Neural Information Processing Systems (NIPS)* 19 (2006) (cit. on p. 65).
- [115] R. Hofstetter, M. Slomczykowski, M. Sati, and L.-P. Nolte. “Fluoroscopy as an imaging means for computer-assisted surgical navigation”. In: *Computer Aided Surgery* 4.2 (1999), pp. 65–76 (cit. on p. 32).
- [116] A. Holzinger and I. Jurisica. “Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions”. In: *Interactive knowledge discovery and data mining in biomedical informatics*. Springer, 2014, pp. 1–18.
- [117] J. Howe. “The rise of crowdsourcing”. In: *Wired Magazine* 14.6 (2006), pp. 1–4 (cit. on p. 72).
- [118] P. J. Huber et al. “Robust estimation of a location parameter”. In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101.
- [119] O. Inel, K. Khamkham, T. Cristea, et al. “CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data”. In: *Proc. Inter. Semantic Web Conf., Part II*. 2014, pp. 486–504 (cit. on p. 72).
- [120] S. Jaeger, A. Karargyris, S. Candemir, et al. “Automatic tuberculosis screening using chest radiographs”. In: *IEEE TMI* 33.2 (2014), pp. 233–245.
- [121] J. M. Jerez, I. Molina, P. J. García-Laencina, et al. “Missing data imputation using statistical and machine learning methods in a real breast cancer problem”. In: *Artificial intelligence in medicine* 50.2 (2010), pp. 105–115 (cit. on p. 9).
- [122] Z. Jiang, Z. Lin, and L. Davis. “Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition”. In: *IEEE Trans Pattern Anal* 35.11 (2013), pp. 2651–2664.
- [123] S. Jonić, C. Sorzano, and N. Boisset. “Comparison of single-particle analysis and electron tomography approaches: an overview”. In: *Journal of microscopy* 232.3 (2008), pp. 562–579 (cit. on p. 23).
- [124] M. I. Jordan and T. M. Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260. eprint: <http://science.sciencemag.org/content/349/6245/255.full.pdf> (cit. on p. 7).
- [125] J. H. Jørgensen, T. L. Jensen, P. C. Hansen, S. H. Jensen, E. Y. Sidky, and X. Pan. “Accelerated gradient methods for total-variation-based CT image reconstruction”. In: *arXiv preprint arXiv:1105.4002* (2011).
- [126] A. C. Kak and M. Slaney. *Principles of computerized tomographic imaging*. IEEE press, 1988.
- [127] M. Kersten, J. Stewart, N. Troje, and R. Ellis. “Enhancing depth perception in translucent volumes”. In: *IEEE TVCG* 12.5 (2006).
- [128] M. Kersten-Oertel, P. Jannin, and D. L. Collins. “The state of the art of visualization in mixed reality image guided surgery”. In: *Comput Med Imag Grap* 37.2 (2013), pp. 98–112.
- [129] F. Kleemann, G. Voß, and K. Rieder. “Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing”. In: *STI Studies* 4.1 (2008) (cit. on p. 72).
- [130] M. Klüppel, J. Wang, D. Bernecker, P. Fischer, and J. Hornegger. “On feature tracking in X-ray images”. In: *Proc. Bildverarbeitung für die Medizin (BVM)*. Springer, 2014, pp. 132–137.
- [131] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. “Dictionary learning algorithms for sparse representation”. In: *Neural computation* 15.2 (2003), pp. 349–396 (cit. on p. 17).

- [132] A. Krizhevsky and G. Hinton. “Learning multiple layers of features from tiny images”. In: (2009) (cit. on p. 8).
- [133] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2012, pp. 1097–1105 (cit. on pp. 4, 17, 72, 73).
- [134] E. A. Krupinski. “The importance of perception research in medical imaging”. In: *Radiation medicine* 18.6 (2000), pp. 329–334.
- [135] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. Duin. “Limits on the majority vote accuracy in classifier fusion”. In: *Pattern Analysis & Applications* 6.1 (2003), pp. 22–31 (cit. on pp. 75, 80).
- [136] C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin. “Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies”. In: *Medical care* 50 (2012), S82–S101 (cit. on p. 10).
- [137] L. Landweber. “An iteration formula for Fredholm integral equations of the first kind”. In: *American journal of mathematics* 73.3 (1951), pp. 615–624.
- [138] T. D. LaToza, W. B. Towne, C. M. Adriano, and A. van der Hoek. “Microtask programming: Building software with a crowd”. In: *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM. 2014, pp. 43–54.
- [139] K. Lawonn, M. Luz, B. Preim, and C. Hansen. “Illustrative Visualization of Vascular Models for Static 2D Representations”. English. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. 2015, pp. 399–406.
- [140] Y. LeCun, B. Boser, J. S. Denker, et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551 (cit. on p. 17).
- [141] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. “Efficient BackProp”. In: *Neural Networks: Tricks of the Trade*. 2012, pp. 9–48 (cit. on p. 76).
- [142] K. Lee, J. Caverlee, and S. Webb. “The social honeypot project: protecting online communities from spammers”. In: *Proc. Int. Conf. World Wide Web*. 2010, pp. 1139–1140 (cit. on p. 81).
- [143] J. M. Lewin, P. K. Isaacs, V. Vance, and F. J. Larke. “Dual-energy contrast-enhanced digital subtraction mammography: feasibility”. In: *Radiology* 229.1 (2003), pp. 261–268 (cit. on p. 32).
- [144] R. Lewis. “Medical phase contrast x-ray imaging: current status and future prospects”. In: *Physics in medicine and biology* 49.16 (2004), p. 3573.
- [145] D. C. Lindberg. “Alhazen’s Theory of Vision and its Reception in the West”. In: *Isis* 58.3 (1967), pp. 321–341.
- [146] N. Littlestone and M. K. Warmuth. “The weighted majority algorithm”. In: *Information and computation* 108.2 (1994), pp. 212–261.
- [147] F. Liu and L. Yang. “A Novel Cell Detection Method Using Deep Convolutional Neural Network and Maximum-Weight Independent Set”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. 2015, pp. 349–357 (cit. on p. 73).
- [148] J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 2015 (cit. on p. 73).
- [149] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. Yu. “Transfer Sparse Coding for Robust Image Representation”. In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. 2013, pp. 407–414.
- [150] V. Lučić, A. Rigort, and W. Baumeister. “Cryo-electron tomography: The challenge of doing structural biology in situ”. In: *Journal of Cell Biol.* 202.3 (Aug. 2013), pp. 407–419 (cit. on p. 62).
- [151] V. Lucic, F. Förster, and W. Baumeister. “Structural studies by electron tomography: from cells to molecules”. In: *Annu. Rev. Biochem.* 74 (2005), pp. 833–865 (cit. on p. 23).

- [152] M. A. Luengo-Oroz, A. Arranz, and J. Frean. “Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears”. In: *Journal of medical Internet research* 14.6 (2012), e167.
- [153] M. Macenko, M. Niethammer, J. S. Marron, et al. “A method for normalizing histology slides for quantitative analysis”. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009, pp. 1107–1110 (cit. on p. 29).
- [154] M. Macenko, M. Niethammer, J. Marron, et al. “A Method for Normalizing Histology Slides for Quantitative Analysis”. In: *ISBI*. Vol. 9. 2009, pp. 1107–1110 (cit. on p. 78).
- [155] L. Maier-Hein, T. Ross, B. Glocker, et al. “Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence”. In: ().
- [156] L. Maier-Hein, S. Mersmann, D. Kondermann, et al. “Crowdsourcing for reference correspondence generation in endoscopic images”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2014*. 2014, pp. 349–356 (cit. on pp. 72, 73).
- [157] L. Maier-Hein, D. Kondermann, T. Roß, et al. “Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences”. In: *Int J Comp Assist Radiol Surg* 10.8 (2015), pp. 1201–1212.
- [158] L. Maier-Hein, P. Mountney, A. Bartoli, et al. “Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery”. In: *Medical image analysis* 17.8 (2013), pp. 974–996 (cit. on p. 33).
- [159] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. “Online dictionary learning for sparse coding”. In: *Proc. Annual International Conference on Machine Learning*. 2009, pp. 689–696 (cit. on p. 17).
- [160] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. “Online learning for matrix factorization and sparse coding”. In: *The Journal of Machine Learning Research* 11 (2010), pp. 19–60.
- [161] J. Mairal, M. Elad, and G. Sapiro. “Sparse representation for color image restoration”. In: *IEEE Transactions on image processing* 17.1 (2008), pp. 53–69 (cit. on p. 17).
- [162] S. G. Mallat and Z. Zhang. “Matching pursuits with time-frequency dictionaries”. In: *IEEE Trans Signal Proc* 41.12 (1993), pp. 3397–3415.
- [163] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš. “A review of 3D/2D registration methods for image-guided interventions”. In: *Medical Image Analysis* 16.3 (2012). Computer Assisted Interventions, pp. 642–661.
- [164] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš. “A review of 3D/2D registration methods for image-guided interventions”. In: *Med Image Anal* 16.3 (2012), pp. 642–661.
- [165] S. Mavandadi, S. Feng, F. Yu, S. Dimitrov, R. Yu, and A. Ozcan. “BioGames: a platform for crowd-sourced biomedical image analysis and telediagnosis”. In: *Games Health J*. 1.5 (2012), pp. 373–376 (cit. on p. 72).
- [166] S. Mavandadi, S. Dimitrov, S. Feng, et al. “Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study”. In: *PloS one* 7.5 (2012), pp. 1932–6203.
- [167] S. Mavandadi, S. Dimitrov, S. Feng, et al. “Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study”. In: *PloS one* 7.5 (2012), e37245.
- [168] V. Mayer-Schönberger and K. Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013 (cit. on p. 3).
- [169] R. B. Mazess, H. S. Barden, J. P. Bisek, and J. Hanson. “Dual-energy x-ray absorptiometry for total-body and regional bone-mineral and soft-tissue composition.” In: *The American journal of clinical nutrition* 51.6 (1990), pp. 1106–1112 (cit. on p. 32).
- [170] B. F. McEwen and M. Marko. “The emergence of electron tomography as an important tool for investigating cellular ultrastructure”. In: *Journal of Histochemistry & Cytochemistry* 49.5 (2001), pp. 553–563 (cit. on p. 24).

- [171] R. McLaughlin, J. Hipwell, D. J. Hawkes, J. A. Noble, J. V. Byrne, T. C. Cox, et al. “A comparison of a similarity-based and a feature-based 2-D-3-D registration method for neurointerventional use”. In: *IEEE Trans Med Imag* 24.8 (2005), pp. 1058–1066.
- [172] S. Miao, Z. J. Wang, and R. Liao. “A CNN Regression Approach for Real-Time 2D/3D Registration”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1352–1363.
- [173] C. A. Mistretta, A. B. Crummy, and C. M. Strother. “Digital angiography: a perspective.” In: *Radiology* 139.2 (1981), pp. 273–276 (cit. on p. 32).
- [174] U. Mitrovic, Z. Spiclin, B. Likar, and F. Pernus. “3D-2D registration of cerebral angiograms: a method and evaluation on clinical images”. In: *IEEE Trans Med Imag* 32.8 (2013), pp. 1550–1563.
- [175] F. Mourgues, F. Devemay, and E. Coste-Maniere. “3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery”. In: *Proceedings IEEE and ACM International Symposium on Augmented Reality*. 2001, pp. 191–192.
- [176] F. Mourgues, F. Devemay, and E. Coste-Maniere. “3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery”. In: *Augmented Reality, 2001. Proceedings. IEEE and ACM International Symposium on*. IEEE. 2001, pp. 191–192.
- [177] V. Nair and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Proc. Int. Conf. Machine Learning (ICML)*. 2010, pp. 807–814 (cit. on p. 75).
- [178] R. Narasimha, I. Aganj, A. Bennett, et al. “Evaluation of denoising algorithms for biological electron tomography”. In: *Journal of structural biology* 164.1 (Oct. 2008), pp. 7–17 (cit. on pp. 27, 66, 67).
- [179] N. Navab, S. M. Heining, and J. Traub. “Camera augmented mobile C-arm (CAMC): calibration, accuracy study, and clinical application”. In: *IEEE Trans Med Imag* 29.7 (2010).
- [180] N. Navab, S. Wiesner, S. Benhimane, E. Euler, and S. M. Heining. “Visual servoing for intraoperative positioning and repositioning of mobile C-arms”. In: *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2006, pp. 551–560.
- [181] S. Nepal, R. Ranjan, and K.-K. R. Choo. “Trustworthy processing of healthcare big data in hybrid clouds”. In: *IEEE Cloud Computing* 2.2 (2015), pp. 78–84 (cit. on p. 3).
- [182] I. Newton. “A letter of mr. isaac newton, professor of the mathematicks in the university of cambridge; containing his new theory about light and colors: Sent by the author to the publisher from cambridge, febr. 6. 1671/72; in order to be communicated to the r. society”. In: *Philosophical Transactions (1665-1678)* 6 (1972), pp. 3075–3087 (cit. on p. 3).
- [183] A. Ng. *What Artificial Intelligence Can and Can't Do Right Now*. 2016 (cit. on p. 4).
- [184] S Nickell, F Förster, A Linaroudis, et al. “TOM software toolbox: acquisition and analysis for electron tomography”. In: *Journal of Structural Biology* 149.3 (2005) (cit. on p. 66).
- [185] D. P. Noonan, P. Mountney, D. S. Elson, A. Darzi, and G.-Z. Yang. “A stereoscopic fibroscope for camera motion and 3D depth recovery during minimally invasive surgery”. In: *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*. IEEE. 2009, pp. 4463–4468 (cit. on p. 33).
- [186] N. A. Obuchowski. “Receiver operating characteristic curves and their use in radiology”. In: *Radiology* 229.1 (2003), pp. 3–8 (cit. on p. 18).
- [187] C. C. Paige and M. A. Saunders. “LSQR: An algorithm for sparse linear equations and sparse least squares”. In: *ACM Transactions on Mathematical Software (TOMS)* 8.1 (1982), pp. 43–71.
- [188] S. H. Park, J. M. Goo, and C.-H. Jo. “Receiver operating characteristic (ROC) curve: practical review for radiologists”. In: *Korean Journal of Radiology* 5.1 (2004), pp. 11–18 (cit. on p. 18).
- [189] M. Pedersen and J. Y. Hardeberg. “Full-reference image quality metrics: Classification and evaluation”. In: *Foundations and Trends® in Computer Graphics and Vision* 7.1 (2012), pp. 1–80 (cit. on p. 21).
- [190] P. A. Penczek. “Chapter three-resolution measures in molecular electron microscopy”. In: *Methods in enzymology* 482 (2010), pp. 73–100.

- [191] H. Peng, R. Rao, S. Dianat, et al. “Multispectral image denoising with optimized vector bilateral filter”. In: *Image Processing, IEEE Transactions on* 23.1 (2014), pp. 264–273 (cit. on p. 62).
- [192] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, and D. J. Hawkes. “A comparison of similarity measures for use in 2-D-3-D medical image registration”. In: *IEEE Trans Med Imag* 17.4 (1998), pp. 586–595.
- [193] P. Perona and J. Malik. “Scale-space and edge detection using anisotropic diffusion”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12.7 (1990), pp. 629–639 (cit. on pp. 62, 63, 66, 68).
- [194] T. M. Peters. “Image-guidance for surgical procedures”. In: *Physics in medicine and biology* 51.14 (2006), R505 (cit. on p. 33).
- [195] A. Pinto, F. Caranci, L. Romano, G. Carrafiello, P. Fonio, and L. Brunese. “Learning from errors in radiology: a comprehensive review”. In: *Seminars in Ultrasound, CT and MRI*. Vol. 33. 4. Elsevier. 2012, pp. 379–382 (cit. on p. 32).
- [196] N. Quoc Viet Hung, N. Tam, L. Tran, and K. Aberer. “An Evaluation of Aggregation Techniques in Crowdsourcing”. English. In: *Web Information Systems Engineering – WISE 2013*. 2013, pp. 1–15 (cit. on pp. 73, 75).
- [197] J. Radon. “1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten”. In: *Classic papers in modern diagnostic radiology* 5 (2005).
- [198] W. Raghupathi and V. Raghupathi. “Big data analytics in healthcare: promise and potential”. In: *Health Information Science and Systems* 2.1 (2014), p. 1 (cit. on p. 3).
- [199] I. Ramirez, P. Sprechmann, and G. Sapiro. “Classification and clustering via dictionary learning with structured incoherence and shared features”. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE. 2010, pp. 3501–3508 (cit. on p. 17).
- [200] V. C. Raykar and S. Yu. “Ranking annotators for crowdsourced labeling tasks”. In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2011, pp. 1809–1817 (cit. on p. 86).
- [201] V. C. Raykar, S. Yu, L. H. Zhao, et al. “Learning from crowds”. In: *J. Machine Learn. Res.* 11 (2010), pp. 1297–1322 (cit. on pp. 73, 75, 77).
- [202] M. Riesenhuber and T. Poggio. “Hierarchical models of object recognition in cortex”. In: *Nature neuroscience* 2.11 (1999), pp. 1019–1025.
- [203] F. Ritter, C. Hansen, V. Dicken, O. Konrad, B. Preim, and H.-O. Peitgen. “Real-Time Illustration of Vascular Structures”. In: *IEEE Trans Vis Comp Graphics* 12.5 (2006), pp. 877–884.
- [204] C. V. Robinson, A. Sali, and W. Baumeister. “The molecular sociology of the cell”. In: *Nature* 450.7172 (Dec. 2007), pp. 973–982 (cit. on p. 62).
- [205] P. Robinson. “Radiology’s Achilles’ heel: error and variation in the interpretation of the Röntgen image.” In: *The British Journal of Radiology* 70.839 (1997), pp. 1085–1098.
- [206] P. Robinson, D. Wilson, A. Coral, A. Murphy, and P. Verow. “Variation between experienced observers in the interpretation of accident and emergency radiographs.” In: *The British journal of radiology* 72.856 (1999), pp. 323–330.
- [207] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *MICCAI*. Springer. 2015, pp. 234–241.
- [208] M. Roychowdhury. *Histologic grading*. 2016 (cit. on p. 28).
- [209] R. Rubinstein, M. Zibulevsky, and M. Elad. *Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit*. Tech. rep. Technion - Computer Science Department, 2008.
- [210] L. I. Rudin, S. Osher, and E. Fatemi. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: Nonlinear Phenomena* 60.1-4 (1992), pp. 259–268.
- [211] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

- [212] J. A. Sacristán and T. Dilla. “No big data without small data: learning health care systems begin and end with the individual patient”. In: *Journal of evaluation in clinical practice* 21.6 (2015), pp. 1014–1017 (cit. on pp. 10, 11).
- [213] A. Saxena, S. H. Chung, and A. Y. Ng. “3-D Depth Reconstruction from a Single Still Image”. In: *Int J Comput Vision* 76.1 (2007), pp. 53–69.
- [214] A. Saxena, S. H. Chung, and A. Y. Ng. “3-d depth reconstruction from a single still image”. In: *International journal of computer vision* 76.1 (2008), pp. 53–69.
- [215] A. Saxena, M. Sun, and A. Y. Ng. “Make3D: Depth Perception from a Single Still Image.” In: *AAAI*. 2008, pp. 1571–1576.
- [216] A. Saxena, M. Sun, and A. Y. Ng. “Make3d: Learning 3d scene structure from a single still image”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.5 (2009), pp. 824–840 (cit. on p. 33).
- [217] J. Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117 (cit. on p. 73).
- [218] R. Schutt and C. O’Neil. *Doing data science: Straight talk from the frontline*. " O’Reilly Media, Inc.", 2013 (cit. on p. 8).
- [219] V. S. Sheng, F. Provost, and P. G. Ipeirotis. “Get another label? improving data quality and data mining using multiple, noisy labelers”. In: *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. 2008, pp. 614–622 (cit. on p. 73).
- [220] A. Sheshadri and M. Lease. “Square: A benchmark for research on computing crowd consensus”. In: *AAAI Conf. Human Comput. Crowdsourc.* 2013.
- [221] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. “The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains”. In: *IEEE Signal Processing Magazine* (2012) (cit. on p. 63).
- [222] E. Y. Sidky and X. Pan. “Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization”. In: *Physics in medicine and biology* 53.17 (2008), p. 4777 (cit. on p. 26).
- [223] L. N. Smith and M. Elad. “Improving Dictionary Learning: Multiple Dictionary Updates and Coefficient Reuse”. In: *IEEE Signal Processing Letters* 20.1 (2013), pp. 79–82.
- [224] R. J. Smith and R. M. Merchant. “Harnessing the crowd to accelerate molecular medicine research”. In: *Trends in Molecular Medicine* 21.7 (2015), pp. 403–405.
- [225] B. Steffens. *Ibn al-Haytham: first scientist*. Morgan Reynolds Pub., 2007.
- [226] K. T. Stolee, J. Saylor, and T. Lund. “Exploring the Benefits of Using Redundant Responses in Crowdsourced Evaluations”. In: *CrowdSourcing in Software Engineering (CSI-SE), 2015 IEEE/ACM 2nd International Workshop on*. 2015, pp. 38–44.
- [227] D. Stoyanov, A. Darzi, and G. Z. Yang. “A practical approach towards accurate dense 3D depth recovery for robotic laparoscopic surgery”. In: *Computer Aided Surgery* 10.4 (2005), pp. 199–208 (cit. on p. 33).
- [228] D. Stoyanov, A. Darzi, and G. Z. Yang. “A practical approach towards accurate dense 3D depth recovery for robotic laparoscopic surgery”. In: *Comp Aid Surg* 10.4 (2005), pp. 199–208.
- [229] D. Stoyanov, A. Darzi, and G. Z. Yang. “Dense 3D depth recovery for soft tissue deformation during robotically assisted laparoscopic surgery”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2004, pp. 41–48 (cit. on p. 33).
- [230] D. Stoyanov, G. P. Mylonas, M. Lerotic, A. J. Chung, and G.-Z. Yang. “Intra-operative visualizations: Perceptual fidelity and human factors”. In: *Journal of Display Technology* 4.4 (2008), pp. 491–501.
- [231] Y. Sun, X. Wang, and X. Tang. “Deep convolutional network cascade for facial point detection”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013.

- [232] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.
- [233] A.-I. Tikhonov. *Solutions of Ill Posed Problems (Scripta series in mathematics)*. Vh Winston, 1977.
- [234] C. Tomasi and R. Manduchi. “Bilateral filtering for gray and color images”. In: *Computer Vision, 1998. Sixth International Conference on*. 1998, pp. 839–846 (cit. on pp. 62, 68).
- [235] C. Tomasi and T. Kanade. *Detection and tracking of point features*. 1991.
- [236] I. Tomek. “Two modifications of CNN”. In: *IEEE Trans. Systems, Man and Cybernetics* 6 (1976), pp. 769–772 (cit. on p. 9).
- [237] L. A. Torre, R. L. Siegel, E. M. Ward, and A. Jemal. “Global cancer incidence and mortality rates and trends—an update”. In: *Cancer Epidemiology and Prevention Biomarkers* 25.1 (2016), pp. 16–27 (cit. on p. 28).
- [238] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal. “Global cancer statistics, 2012”. In: *CA: a cancer journal for clinicians* 65.2 (2015), pp. 87–108.
- [239] A. Toshev and C. Szegedy. “Deeppose: Human pose estimation via deep neural networks”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.
- [240] T. Tuytelaars and K. Mikolajczyk. “Local invariant feature detectors: a survey”. In: *Foundations and trends® in computer graphics and vision* 3.3 (2008), pp. 177–280.
- [241] N. Uwe; and O. Schenk. *Combinatorial scientific computing*. Boca Raton, Fla. : CRC Press, 2012 (cit. on p. 64).
- [10] A. Vahadane, T. Peng, S. Albarqouni, et al. “Structure-preserved color normalization for histological images”. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE. 2015, pp. 1012–1015 (cit. on p. 29).
- [11] A. Vahadane, T. Peng, A. Sethi, et al. “Structure-preserving color normalization and sparse stain separation for histological images”. In: *IEEE transactions on medical imaging* 35.8 (2016), pp. 1962–1971 (cit. on p. 29).
- [242] A. Vedaldi and K. Lenc. “MatConvNet-Convolutional Neural Networks for MATLAB”. In: *Proc. ACM Int. Conf. Multimedia*. 2015 (cit. on p. 78).
- [243] A. Vedaldi and K. Lenc. “MatConvNet-convolutional neural networks for MATLAB”. In: *arXiv preprint arXiv:1412.4564* (2014).
- [244] A. Vedaldi and K. Lenc. “Matconvnet: Convolutional neural networks for matlab”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 689–692.
- [245] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. “Community-based bayesian aggregation models for crowdsourcing”. In: *Proc. Int. Conf. World Wide Web*. ACM. 2014, pp. 155–164 (cit. on p. 85).
- [246] M. Veta, P. J. Van Diest, S. M. Willems, et al. “Assessment of algorithms for mitosis detection in breast cancer histopathology images”. In: *Medical image analysis* 20.1 (2015), pp. 237–248 (cit. on pp. 30, 78).
- [247] S. Virga, V. Dogeanu, P. Fallavollita, R. Ghotbi, N. Navab, and S. Demirci. “Optimal C-arm Positioning for Aortic Interventions”. English. In: *Proc. Bildverarbeitung für die Medizin (BVM)*. Informatik aktuell. 2015, pp. 53–58.
- [248] D. Volpi, M. H. Sarhan, R. Ghotbi, N. Navab, D. Mateus, and S. Demirci. “Online tracking of interventional devices for endovascular aortic repair”. In: *Int J Comp Assist Radiol Surg* 10.6 (2015), pp. 773–781.
- [249] D. Volpi, M. H. Sarhan, R. Ghotbi, N. Navab, D. Mateus, and S. Demirci. “Online tracking of interventional devices for endovascular aortic repair”. In: *Int. J. Comp. Assist. Radiol. Surg.* 10.6 (2015), pp. 773–781 (cit. on p. 72).
- [250] L. Von Ahn. “Games with a purpose”. In: *Computer* 39.6 (2006), pp. 92–94 (cit. on p. 72).

- [251] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. “recaptcha: Human-based character recognition via web security measures”. In: *Science* 321.5895 (2008), pp. 1465–1468 (cit. on p. 72).
- [252] L. M. Voortman, S. Stallinga, R. H. Schoenmakers, L. J. van Vliet, and B. Rieger. “A fast algorithm for computing and correcting the CTF for tilted, thick specimens in TEM”. In: *Ultramicroscopy* 111.8 (2011), pp. 1029–1036.
- [253] X. Wan, F. Zhang, and Z. Liu. “Modified simultaneous algebraic reconstruction technique and its parallelization in cryo-electron tomography”. In: *Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on*. IEEE. 2009, pp. 384–390 (cit. on p. 27).
- [254] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. “Deep learning for identifying metastatic breast cancer”. In: *arXiv preprint arXiv:1606.05718* (2016).
- [255] G. Wang. “A perspective on deep imaging”. In: *IEEE Access* 4 (2016), pp. 8914–8924 (cit. on p. 4).
- [256] J. Wang, M. Kreiser, L. Wang, N. Navab, and P. Fallavollita. “Augmented depth perception visualization in 2D/3D image fusion”. In: *Computerized Medical Imaging and Graphics* 38.8 (2014), pp. 744–752 (cit. on p. 33).
- [257] J. Wang, M. Kreiser, L. Wang, N. Navab, and P. Fallavollita. “Augmented depth perception visualization in 2D/3D image fusion”. In: *Comput Med Imag Graph* 38.8 (2014), pp. 744–752.
- [258] J. Wang, P. Fallavollita, L. Wang, M. Kreiser, and N. Navab. “Augmented reality during angiography: integration of a virtual mirror for improved 2D/3D visualization”. In: *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*. IEEE. 2012, pp. 257–264 (cit. on p. 33).
- [259] J. Wang, A. Borsdorf, J. Endres, and J. Hornegger. “Depth-Aware Template Tracking for Robust Patient Motion Compensation for Interventional 2-D/3-D Image Fusion”. In: *Proc. IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*. Ed. by IEEE. Seoul, South Korea, 2013.
- [260] J. Wang, C. Riess, A. Borsdorf, B. Heigl, and J. Hornegger. “Sparse Depth Sampling for Interventional 2-D/3-D Overlay: Theoretical Error Analysis and Enhanced Motion Estimation”. In: *Proc. Computer Analysis of Images and Patterns*. York, UK, 2013, pp. 86–93.
- [261] J. Wang, J. Wang, J. Yang, et al. “Locality-constrained linear coding for image classification”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010, pp. 3360–3367.
- [262] L. Wang, P. Fallavollita, R. Zou, X. Chen, S. Weidert, and N. Navab. “Closed-form inverse kinematics for interventional C-arm X-ray imaging with six degrees of freedom: modeling and application”. In: *IEEE Trans Med Imag* 31.5 (2012), pp. 1086–1099.
- [263] S. Wang, S. Fidler, and R. Urtasun. “Holistic 3d scene understanding from a single geo-tagged image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3964–3972 (cit. on p. 33).
- [264] X. Wang, C. S. zu Berge, S. Demirci, P. Fallavollita, and N. Navab. “Improved interventional X-ray appearance”. In: *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*. IEEE. 2014, pp. 237–242 (cit. on p. 33).
- [265] X. Wang, C. Schulte zu Berge, S. Demirci, P. Fallavollita, and N. Navab. “Improved Interventional X-ray Appearance”. In: *Proc. ISMAR*. 2014, pp. 237–242.
- [266] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612 (cit. on p. 21).
- [267] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *Image Processing, IEEE Transactions on* 13.4 (2004), pp. 600–612.
- [268] G. Warmerdam, P. Steininger, M. Neuner, G. Sharp, and B. Winey. “Influence of imaging source and panel position uncertainties on the accuracy of 2D/3D image registration of cranial images”. In: *Med Phys* 39.9 (2012).

- [269] Y. Weiss, R. Fergus, and A. Torralba. “Multidimensional Spectral Hashing”. In: *ECCV (5)*. Ed. by A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid. Vol. 7576. LNCS. Springer, 2012, pp. 340–353.
- [270] M. Wertheimer. “Laws of organization in perceptual forms.” In: *A source book of Gestalt psychology (1938)*, pp. 71–88.
- [271] M. Wertheimer. “Untersuchungen zur Lehre von der Gestalt. II”. In: *Psychologische Forschung* 4.1 (1923), pp. 301–350.
- [272] A. E. Weston, H. E. Armer, and L. M. Collinson. “Towards native-state imaging in biological context in the electron microscope”. In: *Journal of chemical biology* 3.3 (2010), pp. 101–112 (cit. on p. 24).
- [273] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. “Whose vote should count more: Optimal integration of labels from labelers of unknown expertise”. In: *Advances in neural information processing systems*. 2009, pp. 2035–2043 (cit. on p. 80).
- [274] M. Wiecezorek, A. Aichert, P. Fallavollita, et al. “Interactive 3D visualization of a single-view X-Ray image”. In: *MICCAI*. Springer. 2011, pp. 73–80 (cit. on p. 33).
- [275] D. Wolf, A. Lubk, and H. Lichte. “Weighted simultaneous iterative reconstruction technique for single-axis tomography”. In: *Ultramicroscopy* 136 (2014), pp. 15–25 (cit. on p. 27).
- [276] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang. “Sparse Multi-Modal Hashing”. In: *IEEE Trans Multimedia* 16.2 (2014), pp. 427–439.
- [277] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu. “MILCut: A Sweeping Line Multiple Instance Learning Paradigm for Interactive Image Segmentation”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 256–263.
- [278] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang. “Beyond Classification: Structured Regression For Robust Cell Detection Using Convolutional Neural Network”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. 2015, pp. 358–365 (cit. on p. 73).
- [279] Q. Xiong, M. K. Morphew, C. L. Schwartz, A. H. Hoenger, and D. N. Mastronarde. “CTF determination and correction for low dose tomographic tilt series”. In: *Journal of structural biology* 168.3 (2009), pp. 378–387 (cit. on p. 25).
- [280] B. Yu, M. Willis, P. Sun, and J. Wang. “Crowdsourcing participatory evaluation of medical pictograms using Amazon Mechanical Turk”. In: *J. Med. Internet Res.* 15.6 (2013), e108 (cit. on p. 72).
- [281] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. “Deconvolutional networks”. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE. 2010, pp. 2528–2535.
- [282] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. “Part-based R-CNNs for fine-grained category detection”. In: *ECCV*. 2014.
- [283] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. “Loss Functions for Neural Networks for Image Processing”. In: *arXiv preprint arXiv:1511.08861* (2015).
- [284] M. Zheng, J. Bu, C. Chen, et al. “Graph Regularized Sparse Coding for Image Representation”. In: *Image Processing, IEEE Transactions on* 20.5 (2011), pp. 1327–1336.
- [285] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu. “3D Deep Learning for Efficient and Robust Landmark Detection in Volumetric Data”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Springer, 2015, pp. 565–572.
- [286] Y. Zheng, B. Georgescu, H. Ling, S. Zhou, M. Scheuering, and D. Comaniciu. “Constrained marginal space learning for efficient 3D anatomical structure detection in medical images”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 194–201.
- [287] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu. “Four-Chamber Heart Modeling and Automatic Segmentation for 3-D Cardiac CT Volumes Using Marginal Space Learning and Steerable Features”. In: *Medical Imaging, IEEE Transactions on* 27.11 (2008), pp. 1668–1681.

- [288] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos. “Machine Learning on Big Data: Opportunities and Challenges”. In: *Neurocomputing* (2017) (cit. on pp. 7, 8).
- [289] S.-A. Zhou and A. Brahme. “Development of phase-contrast X-ray imaging techniques and potential medical applications”. In: *Physica Medica* 24.3 (2008), pp. 129–148 (cit. on p. 32).
- [290] H. Zou and T. Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.
- [291] M. H. Zweig and G. Campbell. “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.” In: *Clinical chemistry* 39.4 (1993), pp. 561–577 (cit. on p. 18).

Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction

Shadi Albarqouni^{*†}, Tobias Lasser[†], Weam Alkhaldi^{*}, Ashraf Al-Amoudi^{*}, and Nassir Navab^{†*}

^{*} Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

^{*} Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Reprint Denied. The reprint of this publication was rejected on open-access platforms. The publication can be found at https://link.springer.com/chapter/10.1007/978-3-319-18431-9_5. The details are provided below.

Copyright Statement. ©2014 Springer and Computational Methods for Molecular Imaging - CMMI 2014, Lecture Notes in Computational Vision and Biomechanics Volume 22, 2014, pp 43-51, Shadi Albarqouni, Tobias Lasser, Weam Alkhaldi, Ashraf Al-Amoudi, Nassir Navab, 'Gradient Projection for Regularized Cryo-Electron Tomographic Reconstruction'. Reprint denied.

Contributions. The author of this thesis was responsible for the main idea of incorporating the Huber regularization in the iterative reconstruction method for Cryo-Electron Tomographic data, and for the validation and testing as well as for writing the manuscript. Co-authors contributed to the background of the iterative reconstruction algorithms, as well as for the revision of the manuscript.

Abstract. Cryo-ET has recently emerged as a leading technique to investigate the three-dimensional (3D) structure of biological specimens at close-to-native state. The technique consists of acquiring many two-dimensional (2D) projections of the structure under scrutiny at various tilt angles under cryogenic conditions. The 3D structure is recovered through a number of steps including projection alignment and reconstruction. However, the resolution currently achieved by cryo-ET is well below the instrumental resolution mainly due to the contrast transfer function of the microscope, the limited tilt range and the high noise power. These limitations make the 3D reconstruction procedure very challenging. Here, we propose a new regularized reconstruction technique based on projected gradient algorithm. Using the gold-standard method for resolution assessment, the Fourier Shell Correlation, we show that the proposed technique outperforms the commonly used reconstruction methods in ET, including the filtered back projection and the algebraic reconstruction techniques.

Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data

Shadi Albarqouni^{*†}, Maximilian Baust[†], Sailesh Conjeti[†], Ashraf Al-Amoudi^{*}, and Nassir Navab^{†*}

^{*} Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

^{*} Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Copyright Statement. ©2015 British Machine Vision Conference (BMVC), 2015, pp 17.1-17.10, Shadi Albarqouni, Maximilian Baust, Sailesh Conjeti, Ashraf Al-Amoudi, Nassir Navab, 'Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data': With kind permission from BMVC.

Contributions. The author of this thesis was responsible for the main idea of incorporating the Graph-Based regularization into the denoising algorithm for Cryo-Electron Tomographics data, and for the validation and testing as well as for writing the manuscript. Co-authors contributed to the revision of the manuscript.

Abstract. Cryo-Electron Tomography is a leading imaging technique in structural biology, which is capable of acquiring two-dimensional projections of cellular structures at high resolution and close-to-native state. Due to the limited electron dose the resulting projections exhibit extremely low SNR and contrast. The 3D structure is then reconstructed and passed through a number of post-processing steps including de-noising and sub-tomogram averaging to provide a better understanding and interpretation. As CET is mainly used for imaging fine scale structures, any denoising method applied to CET images should be scale selective and in particular be able to preserve such fine scale structures. In this context, we propose a new denoising framework based on regularized graph spectral filtering with a full control of scale-space and global consistency. Using the gold-standard metrics, we show that our denoising algorithm significantly outperforms the state-of-the-art methods such as NAD, NLM and RGF in terms of noise removal and structure preservation.

B.1 Introduction

Cryo-electron tomography (CET) is a powerful imaging technique in biological sciences which bridges the gap between the molecular and the cellular structural biology [204], giving a better understanding of protein interactions and thus better drug delivery strategies. In principle, similar to Computed Tomography (CT) in Medical Imaging, CET acquires two-dimensional projections at high resolution (around 20-50 Angstrom) of three dimensional (3D) cellular structures (called tomograms) at cryogenic (freezing) conditions under near-to-native state. Due to the low electron dose, necessary to avoid biological specimen damage, and limited tilt angle (typically $\pm 60^\circ$ to 70°), a noisy (SNR typically 0.1 to 0.01) and extremely weak signal (low contrast) is formed in the resulting projections. These unfiltered projection images are then projected back to build the tomogram. In the reconstruction phase, the noise is propagated through the tomogram making the noise model more complicated. The reader is referred to [80] for more details on image formation and noise model in Electron Microscopy.

Therefore, post-processing steps, such as noise reduction, after 3D reconstruction are necessary to provide a better visualisation and interpretation of the structure under scrutiny. However, this process is critical and could lead to wrong interpretation by erroneously removing fine structural information, that can not be discriminated from the noise. Conventional linear filters such as Gaussian kernels succeed in reducing noise, however, at the expense of blurring edges. Popular Non-linear anisotropic diffusion (NAD) [193] and its extended versions, which can be interpreted in terms of scale space theory, are extensively used in CET community due to their successful performance. However, NAD requires the diffusivity to be chosen carefully, which is sometimes quite challenging, and needs many iterations to converge. Non-local means (NLM) filter and its fast version [50] are investigated for denoising tomograms having redundant information, however, their performance is degraded when the window size is increased (spatial neighbourhood), in particular for high resolution CET data ($> 2048^2$). To date, both techniques are still used in the context of CET, however, advanced filtering algorithms which are able to smooth the noise while preserving the edges to increase the contrast as well are still in demand and highly desirable [150].

Bilateral filter (BF) [234] and its vectorized extension [191] have been successfully applied in computer vision community. One related technique is the rolling guidance filter (RGF) [77], which can be interpreted in terms of joint bilateral filter. However, it uses the filtered image as guidance rather than the original image which is commonly used in guided filters. This way, it succeeds in preserving the edges while smoothing the background. Another related work is [85], where bilateral filtering (BF) is

interpreted in graph spectral domain addressing some open issues in [221] regarding emerging signal processing on graphs. As mentioned before, denoising CET images requires a proper scale selection as well as the preservation of fine scale structures. The proposed method is thus based on the following considerations:

- By using a multi-scale pyramid for guidance we are able to detect meaningful scales and use them for guidance without oversmoothing fine scale structures.
- Using a patch-based approach, we can take advantage of redundant structures in the whole image rather than using a pre-defined spatial window for averaging similar pixels or patches. This way, we can preserve the local and global consistencies.
- By deriving explicit solution formulas for computing the intermediate filtering results we obtain an efficient algorithm.

Inspired by [77] and [85], we propose the *Multi-scale Graph-based Guided Filter (MG²F)*, which is - to the best of our knowledge - the first attempt of employing multi-scale graph representation as a guidance for an iterative graph spectral filtering in general and on CET data in particular.

B.2 Methodology

We assume the noisy image I_η to be corrupted by white Gaussian noise, thus a suitable objective function would be

$$\hat{I}_f = \arg \min_{I_f} \frac{1}{2} \|I_f - I_\eta\|_2^2, \quad (\text{B.1})$$

but we will augment this energy by a novel multi-scale graph regularisation as described in the following.

B.2.1 Graph Representation

Given a noisy image I_η , we collect N overlapping patches covering the whole image $P \in \mathbb{R}^{\sqrt{n} \times \sqrt{n}}$, which can be seen as data points $\nu = \{\nu_1, \nu_2, \dots, \nu_N\} \in \mathbb{R}^{n \times N}$ lying on a manifold \mathcal{M} embedded in \mathbb{R}^n space such that $\nu = EI_\eta$, where E is an operator collecting patches and vectorize it, cf. Figure B.1. The relation between the data points can be represented by a k -NN connected, undirected, and weighted graph $G = \{\nu, \varepsilon, \omega\}$, where ν is the data points (patches), ε is the set of edges, and ω is the set of edge weights.

Weight Assignment

Assigning weights to the edges which exhibit a low SNR such in Cryo-ET data is challenging, therefore, we recall the scale-space theory [193] to build a Gaussian pyramid $I_{G_{\sigma_s}} = G_{\sigma_s} * I_\eta$ such that the noise manifests itself at certain structure scale σ_s and the semantical image appears clearly as shown in Figure B.1, then the weights of the data points can be easily assigned using a heat kernel as follows:

$$W_{ij} = \begin{cases} \exp - \frac{\|\nu_i - \nu_j\|_{2, \sigma_s}^2}{\sigma_h^2}, & \varepsilon_{ij} \in k\text{-NN}, \\ 0, & \textit{else}. \end{cases} \quad (\text{B.2})$$

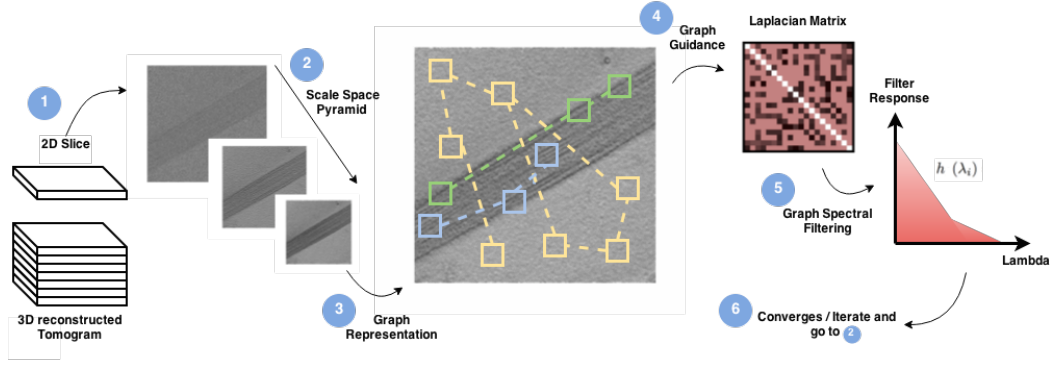


Figure B.1. MG²F Framework: A noisy image slice from the 3D reconstructed tomogram is fed to the algorithm, where the graph is built on a selected scale space image (i.e. coarse grid) acting as a guidance for the regularized graph spectral filter.

Where $\|\cdot\|_{2,\sigma}$ is the Euclidean distance between two vectors at scale σ , however, σ_h is controlling the affinity of the neighbouring data points. Further, we denote the diagonal degree matrix by D , where $D_{ii} := \sum_j W_{ij}$.

Graph Guidance Regularization

In this work, we are interested to preserve the intrinsic structure of the data points v in the spectral filtering phase. It is worth mentioning that v will be collected from the iterated filtered image I_f . We recall the definition of the Laplacian Quadratic Form [241], which can be represented as follows:

$$S_{\sigma_s}(I_f) = \sum_{(i,j) \in \mathcal{E}} W_{ij} \|v_i - v_j\|_2^2 = \frac{1}{2} \text{Tr}(v L_{\sigma_s} v^T). \quad (\text{B.3})$$

This expression can be interpreted as a regularization term that minimizes the distance between data points guided by, we denote it, the penalty Laplacian graph $L_{\sigma_s} := D - W$, which computed at different scales σ_s . The normalised Laplacian graph can be computed by $\tilde{L}_{\sigma_s} := D^{-\frac{1}{2}} L_{\sigma_s} D^{-\frac{1}{2}}$. The reader is referred to [95] for more details on graph operators.

B.2.2 Graph Spectral Filtering

The spectrum of the graph $\sigma(G)$ can be obtained from the eigenvalue decomposition of the normalised graph Laplacian $\tilde{L}_{\sigma_s} := U \Lambda U^T$, where the eigenvalues $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_N\} \in [0, 1]$ carry a notion of the graph frequencies, and the eigenvectors $U^T := \{u_1, u_2, \dots, u_N\}^T \in \mathbb{R}^{N \times n}$ act as the orthogonal basis of the Graph Fourier Transform (GFT) [95], so we can write the transformed signal as follows $\hat{v} = U^T v$.

Regularized Energy

We define our objective function as follows:

$$\hat{I}_f = \arg \min_{I_f} \left\{ \frac{1}{2} \|I_f - I_\eta\|_2^2 + \alpha S_{\sigma_s}(I_f) \right\}, \quad (\text{B.4})$$

where $\alpha > 0$ is the regularization parameter. The solution can be written in a closed form:

$$\hat{I}_f = E^T \left(\sum_{i=1}^N \frac{1}{(1 + \alpha\lambda_i)} u_i \hat{v}_i \right) = E^T \left(\frac{1}{I + \alpha \tilde{L}_{\sigma_s}} \right) E I_\eta, \quad (\text{B.5})$$

where E^T denotes the reshaping process of the previously vectorised patches. It becomes apparent from (B.5) that the signal is filtered on the spectral domain before doing the inverse GFT, where the spectral response of the filter $h_1(\lambda_i) = 1/(1 + \alpha\lambda_i)$ controls the frequency decay and thus the degree of smoothness.

B.2.3 Connection to Classical Filters

Different classical filters can be expressed similar to (B.5) with different spectral filters, for instance, the Bilateral filter (BF) kernel can be written as $\nu_{BF} = D^{-1} W \nu$, where its spectral response can be recast as a linear spectral filter $h(\lambda_i) = (1 - \lambda_i)$ [85], the same applies for non-local means filter (NLM), while the nonlinear anisotropic diffusion (NAD) has an exponential spectral filter $h_2(\lambda_i) = e^{-\alpha\lambda_i}$. For the sake of having the power of diffusivity (fast decay) along with the regularized graph, we propose a new spectral filter

$$h_3(\lambda_i) = e^{-\kappa\alpha\lambda_i} / (1 + \alpha\lambda_i), \quad (\text{B.6})$$

where κ is a decaying factor. The proposed filter can be interpreted in the context of fractional derivative orders of Laplacian in Sobolev space [56], that shows a promising performance. A comparison of different filter responses is shown in Figure B.2.

B.2.4 Stopping Criterion

One can simply raise a question, why we need a stopping criterion where we have already a closed form solution for (B.4). Indeed, this optimal solution is designated for a specific scale, and since we are interested in having a multi-scale reconstruction, the resultant filtered image from the previous scale used as a guidance for the next scale, hence the need of a stopping criterion. Choosing it automatically is an important feature for variational approaches in general, [114] suggested one stopping criterion for Manifold de-noising, based on graph diffusion, therefore we employ the graph diffusion distance proposed by [107], $\xi(L^k, L^{k-1}) := \|e^{-L^k} - e^{-L^{k-1}}\|_F^2$, which computes each iteration the distance between consequent graph Laplacians, which reflects the significant change in the filtering process. Then the optimization problem formulated as follows:

$$\hat{I}_f = \arg \min_{I_f, \sigma_s} \left\{ \frac{1}{2} \|I_f - I_\eta\|_2^2 + \alpha S_{\sigma_s}(I_f) \right\}, \quad \text{s.t.} \quad \xi(L^k, L^{k-1}) \leq \beta, \quad (\text{B.7})$$

where β is the desired distance and k is the iteration index, which can be minimized by Algorithm 1.

B.3 Experiments and Results

Our experiments are conducted on computer vision, simulated data, that rather mimics the complicated noise model in CET, as well as real CET tomographic data. We compare the results of our algorithm (MG²F) against common de-noising gold-standard filters in computer vision community, then we

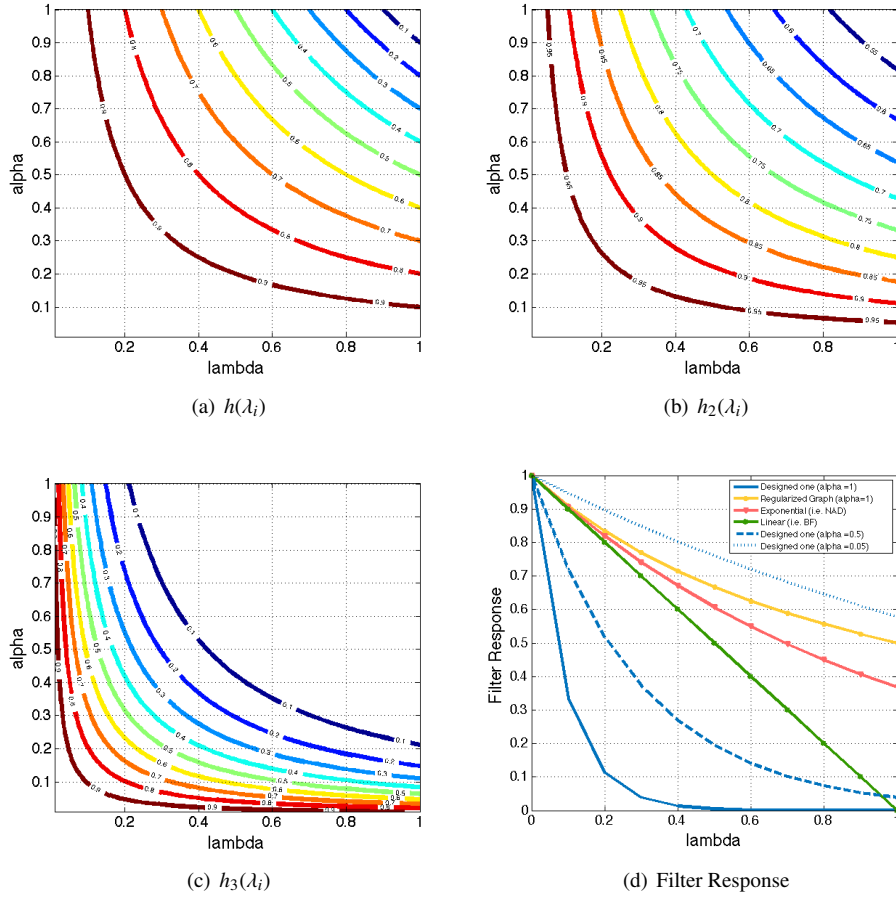


Figure B.2. Spectral filters responses: (a) linear (i.e. Bilateral), (b) regularised graph, and (c) designed one, against the parameters λ (spectral frequency) and α (regularization parameter), (d) shows the line profile ($\alpha = 1$) for different filters.

compare it with the successful filters in CET [178] such as Nonlinear Anisotropic Diffusion (NAD) [193] and Fast Non-Local Means (NLM) [50], and further with the recent scale-aware filter, the Rolling Guidance Filter (RGF) [77]. The filter parameters are tuned in an optimal way; either from the cited references or determined experimentally in order to visualise the feature of interest. Results are validated by different metrics; i.e. data with ground truth are validated using Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE), however, we followed [80] and [178] for evaluating denoising methods on real data.

Computer Vision: To give a good illustrative example, we run the algorithm on Lena image, which corrupted by an (i.i.d) Gaussian noise resulting in SNR of 7. Different algorithms are applied on this image, results are shown in Figure B.3 for the cropped images. It is clear that our method gives an outperforming PSNR indicating for better contrast.

Simulated Data: A GroEL tomogram, which is obtained from the Electron Microscopy Data Bank (EMDB-ID: 1AON), is generated using the TOM toolbox [184], where both (i.i.d) Gaussian and Poisson noise are added on the projection images resulting in a signal-to-noise ratio (SNR) of 0.1, then a slice of the reconstructed tomogram (the noisy image) is passed to different denoising algorithms. We run the algorithms on 150 random slices collected from 15 different tomograms. Results are

Require: The noisy image I_η , patch size \sqrt{n} , σ_s , k -NN, k_{max} , α and β .

Ensure: The filtered image I_f in (B.7).

- 1: Initialize the Laplacian $L^0 = ones(N, N)$, $I_f^0 \leftarrow I_\eta$
- 2: **while** $\xi(L^k, L^{k-1}) > \beta$ **or** $k < k_{max}$ **do**
- 3: Find the scale-space image $I_{G_{\sigma_s}} = G_{\sigma_s} * I_f^k$.
- 4: $\sigma_s \leftarrow$ next finer level in the pyramid.
- 5: Collect patches (data points) $v = EI_f^k$.
- 6: Build the graph & assign weights for the k -NN patches $v_{\sigma_s} = EI_{G_{\sigma_s}}$ using (B.2).
- 7: Compute the normalised graph Laplacian \tilde{L}_{σ_s} and the graph spectrum $\sigma(G)$.
- 8: Apply the spectral filtering $I_f^{k+1} = E^T \left(\sum_{i=1}^N h(\lambda_i) u_i \hat{v}_i \right)$.
- 9: Compute the graph diffusion $\xi(L^k, L^{k-1})$.
- 10: $I_f^k \leftarrow I_f^{k+1}$
- 11: **end while**

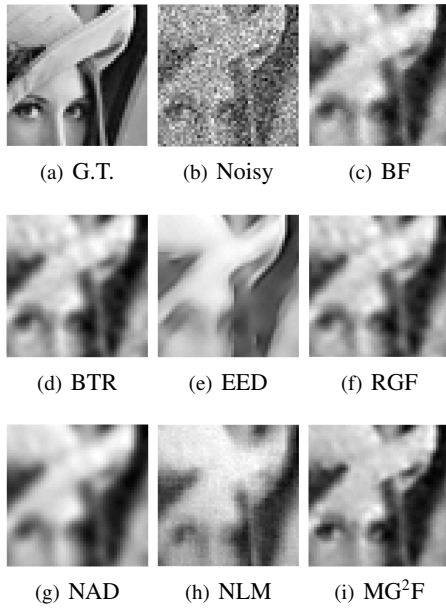
Algorithm 1: MG²F algorithm

validated by PSNR, our method shows significant performance ($p < 0.01$ by t-test, and $p < 0.05$ by Kolmogorov-Smirnov test), making the results consistently better compared to the other methods as shown in Figure B.4(c).

Real CET Data: We also denoised an unstained CET HIV-1 virion (EMDB-ID: 1155), which can be considered as a benchmark data for de-noising in CET. Results are validated by the common validation measure in CET community for 2D images, namely, the Contrast to Noise ratio (CNR) [80]. Further, in the qualitative evaluation session with our clinical partners, they appreciated the enhanced contrast in our method, because the background in Figure B.5(e) was carefully smoothed, while spiky signals such as membrane proteins (arrows) as well as the inner core of the HIV virus (ellipse) were well preserved. As the algorithm takes factors, such as patch size, redundancy, and scale-space, into account, we may conclude, based on their feedback and the resultant CNRs, that it is well suited for handling CET data.

3D Extension: Showing 3D data augments the computed statistics visually and gives a better interpretation for scientists, therefore, we extended the algorithm to handle 3D objects, which conceptually similar to our basic algorithm, however, blocks are collected from volumes rather than patches. An unstained HIV-1 tomogram is denoised using NAD (commonly used in CET), BM3D [49] (commonly used in Computer Vision) and our algorithm *MG²F*. The results are validated using both KL-divergence test ($p < 0.1$) and Fourier Shell Correlation (FSC) used in [178] as shown in Figure B.6. The FSC curve shows the correlation of the corresponding frequency shells between the unprocessed/noisy and denoised tomograms, the blue line shows the auto-correlation of noisy data over the frequency, NAD has a smooth curve as expected due to the diffusion effect, BM3D is slightly better than MG²F in low frequencies, however, MG²F has a sharp decay in higher frequencies which is not the case for BM3D as shown in Figure B.4(d). It is worth mentioning that the higher 0.5-cut-off frequency the higher resolution you get for these tomograms. Therefore, we can say MG²F performs better than NAD (has lower resolution), and BM3D (pass the higher frequencies).

Sensitivity Analysis: Cross validation in general is a daunting task, however, it becomes more difficult when the feature assessment depends mainly on the experts opinion. Therefore, we performed a sensitivity analysis to investigate the effect of these hyperparameters and suggest to update the scale-space parameter σ_s iteratively from the coarse level (given) to the fine level (until saturated). σ_h should be set based on the variance of data points, however, for the sake of simplicity, we normalize the data points before computing the weights. The impact of selecting different k -NN and Patch sizes



Algorithm	PSNR	MSE
Parameters	(dB)	(10^{-3})
Bilateral (BF) [234] <small>($\sigma_l=0.5, \sigma_r=1.5, W=10$)</small>	17.49	174
Beltrami (BTR) [74] <small>($\delta=0.1, iter=10$)</small>	17.37	176
EED [104] <small>($\rho=4, iter=30$)</small>	11.27	324
NAD [193] <small>($iter=10, \kappa=0.3$)</small>	16.50	192
NLM [50] <small>($P=7, W=21, \sigma_S=4\sigma_N$)</small>	12.11	298
RGF [77] <small>($\sigma_l=0.5, \sigma_r=1.5, iter=10$)</small>	17.49	174
MG²F <small>($\alpha=0.8, iter=4, \sigma_h=0.1$)</small>	17.78	169

Figure B.3. Photographic Image: Results of different algorithms on Lena image(128X128, SNR=7) along with a tabulated comparison to the proposed MG²F filter.

(i.e. 3,5,7,9 and 11) on PSNR is shown in Figure B.4. We observe that the algorithm converges at 3-7 iterations demonstrating the performance of the stopping criterion.

B.4 Conclusion

In this chapter, we propose MG²F algorithm for denoising CET data, which incorporates a multi-scale pyramid taking the advantage of redundant structures on different scales into account. This acts as a guidance for the graph spectral filter and this way the local and global consistencies are well preserved. To the best of our knowledge this is the first approach which incorporates a multi-scale scheme in a guided filtering framework. Furthermore, the algorithm converges within only a few iterations and we demonstrated the performance of it on simulated as well as real data.

B.5 References

- [49] A. Danielyan, V. Katkovnik, and K. Egiazarian. “BM3D frames and variational image de-blurring”. In: *Image Processing, IEEE Transactions on* 21.4 (2012), pp. 1715–1728 (cit. on p. 67).
- [50] J. Darbon, A. Cunha, T. Chan, S. Osher, and G. Jensen. “Fast nonlocal filtering applied to electron cryomicroscopy”. In: *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*. 2008, pp. 1331–1334 (cit. on pp. 27, 62, 66, 68).

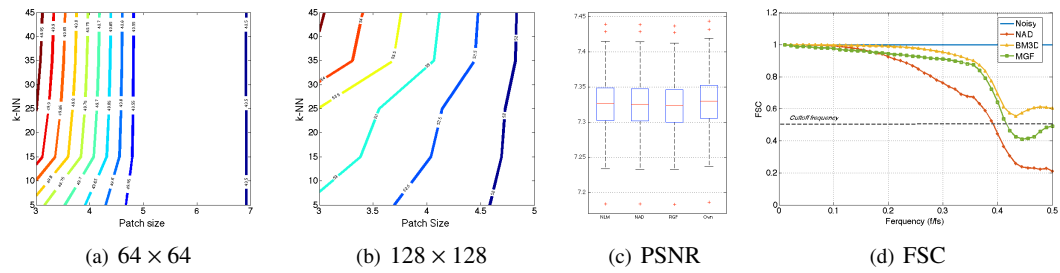


Figure B.4. Sensitivity analysis: PSNR contour against k-NN and Patch size for different image sizes (a) 64^2 and (b) 128^2 . In (c) PSNR of different denoising algorithms (NLM, NAD, RGF and Ours respectively) for 150 slices (SNR=0.1) from 15 simulated tomograms. (d) FSC curve for different denoised 3D tomograms.

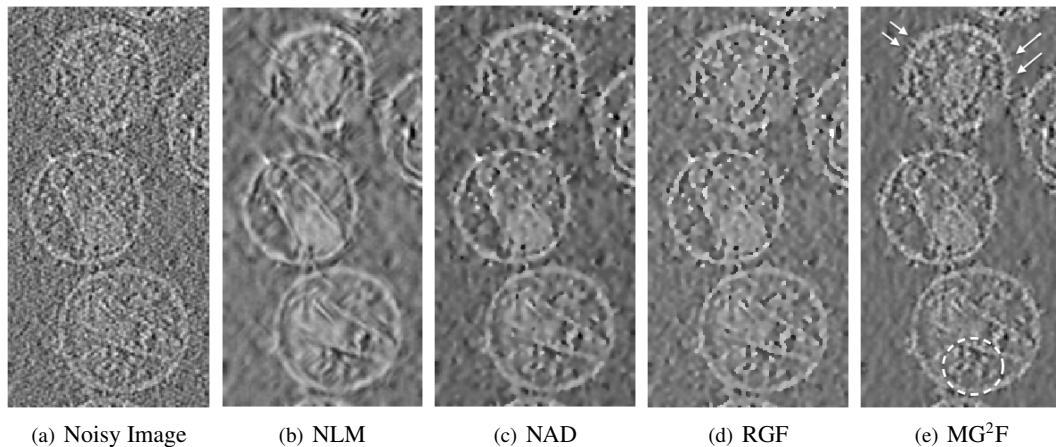


Figure B.5. 2D CET data: Filtering results on the tomogram along with the corresponding CNR of b) NLM (0.1979), c) NAD (0.2570), d) RGF (0.3146), e) Proposed MG²F (**0.3150**), where the arrows point to the fine structures on the membrane and the ellipse contains the inner core of HIV virus.

- [74] J.-J. Fernandez. “TOMOBFLOW: feature-preserving noise filtering for electron tomography”. In: *BMC bioinformatics* 10.1 (2009), p. 178 (cit. on pp. 27, 68).
- [75] J.-J. Fernández and S. Li. “An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms”. In: *Journal of structural biology* 144.1 (2003), pp. 152–161 (cit. on p. 27).
- [95] L. J. Grady and J. R. Polimeni. *Discrete Calculus, Applied Analysis on Graphs for Computational Science*. Springer, 2010 (cit. on p. 64).
- [104] B. ter Haar Romeny, L. Florack, J. Koenderink, M. Viergever, and J. Weickert. “A review of nonlinear diffusion filtering”. In: *Scale-Space Theory in Computer Vision*. Vol. 1252. Springer Berlin Heidelberg, 1997, pp. 1–28 (cit. on p. 68).
- [114] M. Hein and M. Maier. “Manifold denoising”. In: *Advanced in Neural Information Processing Systems (NIPS)* 19 (2006) (cit. on p. 65).
- [150] V. Lučić, A. Rigort, and W. Baumeister. “Cryo-electron tomography: The challenge of doing structural biology in situ”. In: *Journal of Cell Biol.* 202.3 (Aug. 2013), pp. 407–419 (cit. on p. 62).

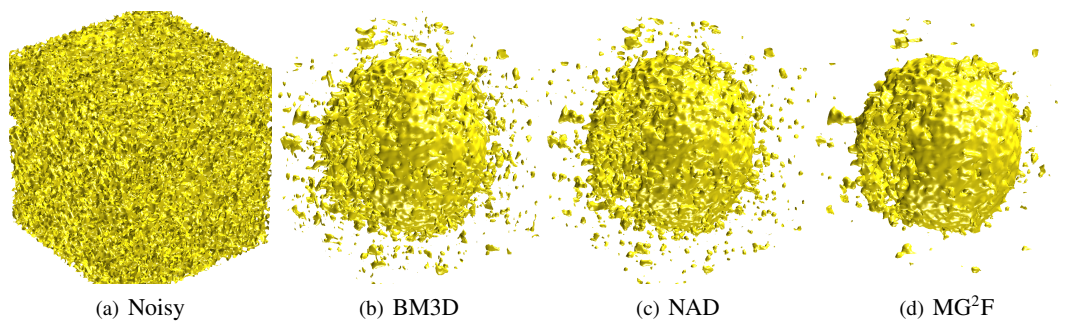


Figure B.6. 3D CET Data: A comparison between different 3D filtering methods to our proposed MG²F method on real unstained HIV-1 data (EMDB-ID: 1155).

- [178] R. Narasimha, I. Aganj, A. Bennett, et al. “Evaluation of denoising algorithms for biological electron tomography”. In: *Journal of structural biology* 164.1 (Oct. 2008), pp. 7–17 (cit. on pp. 27, 66, 67).
- [184] S Nickell, F Förster, A Linaroudis, et al. “TOM software toolbox: acquisition and analysis for electron tomography”. In: *Journal of Structural Biology* 149.3 (2005) (cit. on p. 66).
- [191] H. Peng, R. Rao, S. Dianat, et al. “Multispectral image denoising with optimized vector bilateral filter”. In: *Image Processing, IEEE Transactions on* 23.1 (2014), pp. 264–273 (cit. on p. 62).
- [193] P. Perona and J. Malik. “Scale-space and edge detection using anisotropic diffusion”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12.7 (1990), pp. 629–639 (cit. on pp. 62, 63, 66, 68).
- [221] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. “The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains”. In: *IEEE Signal Processing Magazine* (2012) (cit. on p. 63).
- [241] N. Uwe; and O. Schenk. *Combinatorial scientific computing*. Boca Raton, Fla. : CRC Press, 2012 (cit. on p. 64).

AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images

Shadi Albarqouni^{*†}, Christoph Baur[†], Felix Achilles[†], Vasileios Belagiannis[‡], Stefanie Demirci[†] and Nassir Navab^{†*}

* Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany

† Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

‡ Visual Geometry Group, University of Oxford, Oxford, U.K

* Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Copyright Statement. ©2016 IEEE. Reprinted, with permission, from Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab, 'AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images', IEEE Transaction on Medical Imaging, Special Issue on Deep Learning, May 2016.

Contributions. The author of this thesis was responsible for the main idea of i) having multi-scale CNN and ii) incorporating the Crowdsourcing into deep learning framework for Mitosis Detection in Breast Cancer Histology Images, and for the validation and testing as well as for writing the manuscript. Christoph Baur, joint first author, was responsible for the validation and testing of the multi-scale CNN approach. Co-authors contributed to the revision of the manuscript.

Abstract. The lack of publicly available ground-truth data has been identified as the major challenge for transferring recent developments in deep learning to the biomedical imaging domain. Though crowdsourcing has enabled annotation of large scale databases for real world images, its application for biomedical purposes requires a deeper understanding and hence, more precise definition of the actual annotation task. The fact that expert tasks are being outsourced to non-expert users may lead to noisy annotations introducing disagreement between users. Despite being a valuable resource for learning annotation models from crowdsourcing, conventional machine-learning methods may have difficulties dealing with noisy annotations during training. In this manuscript, we present a new concept for learning from crowds that handle data aggregation directly as part of the learning process of the convolutional neural network (CNN) via additional crowdsourcing layer (*AggNet*). Besides, we present an experimental study on learning from crowds designed to answer the following questions: (i) Can deep CNN be trained with data collected from crowdsourcing?, (ii) How to adapt the CNN to train on multiple types of annotation datasets (ground truth and crowd-based)?, (iii) How does the choice of annotation and aggregation affect the accuracy? Our experimental setup involved Annot8, a self-implemented web-platform based on Crowdfunder API realizing image annotation tasks for a publicly available biomedical image database. Our results give valuable insights into the functionality of deep CNN learning from crowd annotations and prove the necessity of data aggregation integration.

C.1 Introduction

Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization or a company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via flexible open call, the voluntary undertaking of a task [69]. It was first introduced by Jeff Howe and Mark Robinson in 2005 using the internet for outsourcing work to a crowd of people [117]. Being initially considered as market research strategy [129], it is nowadays widely seen as an economical way to recruit crowds for tedious and time-consuming tasks such as annotations for character recognition [251], image classification [133], and natural language processing [119]. As a result of this trend, many crowdsourcing platforms such as Amazon Mechanical Turk (AMT)¹, Games with a Purpose² [250], Crowdfunder³, and LabelMe⁴ have emerged within the past decade. Here, users are not only confronted with simple, every-day tasks, but are also engaged in highly complex processes involving innovation creation.

A good example for this, is the medical domain where, very recently, crowdsourcing has been presented as a solution to the immense lack in publicly available ground-truth data. Various applications such as medical pictogram [280], correspondence finding for stereo endoscopic imaging [156], device detection in angiographic sequences [249], telepathology [165], and medical image segmentation [103] and classification [78] have already shown that crowdsourcing can provide efficient and inexpensive data annotation. With the very recent launch of the CrowdTruth framework⁵, IBM, Google and Amsterdam University have paved the way towards machine-human computing for collecting ground-truth annotation data on text, images and videos in the medical domain. Similarly, Celi *et al.* [38] organised several events and data marathons, where engineers, data scientists, and clinicians were

¹Amazon Mechanical Turk, <https://www.mturk.com>

²Game with a Purpose, <https://www.gwap.com>

³Crowdfunder, <http://www.crowdfunder.com/>

⁴LabelMe, <http://labelme.csail.mit.edu/>

⁵CrowdTruth framework - <http://crowdtruth.org>

invited to address specific challenges during the clinical routines and procedures. As a result, many innovative ideas and prototypes have been developed, clinicians as well as medical students become part of a data-driven learning system. The most astonishing fact about crowdsourcing studies in the medical domain, however, is the conclusion that a crowd of non-professional, inexperienced users do not underperform medical experts [22, 156].

Improving the crowd's quality is very essential for being able to generate a reliable ground-truth and creating an interest within the research community. Redundancy and Aggregation (R&A) (i.e. majority voting) is the baseline approach that has been proposed in this context [196, 219]. However, there is no control on the sensitivity and specificity of single participants. All aforementioned crowdsourcing platforms integrate qualification tests in order to restrict "noisy" annotations. This information can then be incorporated into the ground-truth generation process via aggregation. Recently, Raykar *et al.* [201] have proposed a probabilistic model for supervised learning to evaluate different users and estimate the ground-truth labels. Having such ground-truth is very important for both training many machine learning algorithms as well as for evaluation.

Indeed, deep learning has advanced the field of computer vision the last few years [217] leading to powerful methods for various applications such as object classification [133], detection [87], segmentation [148], robust regression [26] and depth prediction [65]. The most established realization of deep learning are Convolutional Neural Networks (ConvNets or CNN) that have also been successfully applied for biomedical imaging purposes [36, 43, 147, 278]. The bottleneck, however, for deep CNN to yield decent accuracy is the availability of a large number of annotated training samples. In particular in the biomedical domain, sufficient resources are not available.

We believe that crowdsourcing platforms will engage various crowds to collaborate with clinicians and frontline healthcare workers in translating questions into methodologies and innovative solutions of which ground truth data is an essential part. However, it is not clear how state-of-the-art machine learning methods behave when fed with training data consisting of reliable (expert) and unreliable (crowd) annotations [22]. As suggested by Aroyo *et al.* [22], it is our goal to evaluate the trustworthiness of participants and integrate this knowledge into the analysis and further processing of annotations.

In this manuscript, we present a first attempt to apply the concept of learning from crowds within a biomedical environment. Being inspired by prominent previous work in this field [43, 201], we define the specific contribution of our own work as: i) Learning of a multi-scale CNN model for mitosis detection, ii) Incorporation of aggregation schemes into CNN layers, and iii) Augmentation and retraining of the CNN model with crowd's annotation labels.

In our analysis comparing performance of the CNN model when incorporating different types of aggregations schemes, we aim at answering the following questions: i) Can deep CNN be trained with data collected from crowdsourcing and is it robust against "noisy" labels?, ii) How to adapt the CNN when we have both ground-truth label and multiple annotations that could be "noisy"?, and iii) How is the accuracy compared to that obtained by ground-truth or majority voting?

In this manuscript, after recapitulating previous work in this field, we introduce *AggNet*, a novel aggregation layer that is integrated into our multiscale CNN. We further present an analysis of the behavior of CNN with and without aggregation on a publicly available large-scale pathological dataset (including ground truth annotations). However, to the best of our knowledge, there has not yet been

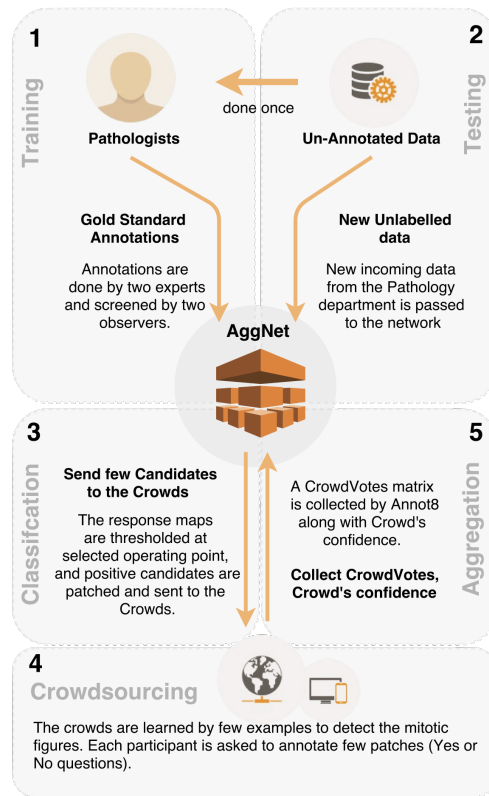


Figure C.1. *AggNet* Framework: (1) The multi-scale CNN model is trained from gold-standard annotations. (2) Then for any incoming unlabelled image, (3) the *AggNet* will produce a response map which is thresholded at selected optimal operating point. (4) These few resulting positive candidates are outsourced to crowds. (5) *AggNet* collects back the crowd votes and jointly aggregates the ground truth and refine the CNN model.

any effort to incorporate this information into machine learning algorithms analyzing the quality of models learned from non-expert annotations.

C.2 Methodology

In this section, we introduce the proposed CNN for aggregating annotations from crowds in conjunction with learning a model for a challenging classification task. Unlike typical supervised methods, which learn a model from ground truth labeled data, learning from crowd annotations is different in the sense that there may be (possibly noisy) multiple labels for the same sample. Our idea is to learn multiple CNN models with the same basic architecture on different image scales (c.f. step 1 in Fig. C.1), perform mitosis detection using these models (c.f. step 2 in Fig. C.1) and provide the crowds with detected mitosis candidates for annotation (c.f. step 3 in Fig. C.1). The collected annotations are then passed to the existing CNN (c.f. step 5 in Fig. C.1) with our aggregation layer attached in order to refine the models and simultaneously generate a ground-truth. This multi-scale approach ensures that we have redundant responses of the same data instances at different scales, with the goal to increase robustness of both aggregation and classification.

C.2.1 Notation

The input to our network is an observation set $D = \{x_i, y_i^j; i = 1, \dots, N, j = 1, \dots, P\}$ containing N instances of $x_i \in \mathbb{R}^d$ (RGB image as d -dimensional vector) with corresponding labels $y_i \in C$ (i.e. $C := \{0, 1\}$ for binary classification) annotated by P independent participants. The goal is to learn a robust CNN model, represented by $f : \mathcal{X} \rightarrow \mathcal{Y}$, from aggregated labels which generalizes well on unseen data:

$$\hat{p} = f(\mathbf{x}, \mathbf{y}; \theta), \quad (\text{C.1})$$

where \hat{p} is the predicted label for an unseen image \mathbf{x} , and θ is the learned model parameter.

C.2.2 Multi-scale CNN Model

Our network architecture consists of three convolutional blocks followed by two fully connected (FC) layers as shown in Fig. C.2. Each convolutional block consists of a convolutional layer followed by a rectified linear unit (ReLU) [177] and max-pooling layer. The output of the softmax layer is the probabilistic score of the mitotic figures.

In our proposed multi-scale CNN model the input image is first down-sampled to different scales (i.e. 0.33, 0.66 and 1). Then, 33×33 patches are collected and passed to the model (scale-wise). On new unlabelled data, we apply the learned model to mirrored and rotated versions (0, 90, 180, and 270 deg) of each image and compute a final detection map (FDM) as the mean of all those detection results.

FDM of different scales are then geometrically averaged to filter out weak responses. By doing this, we aim at obtaining more accurate detections.

During learning from crowd annotations phase, we augment the CNN architecture with our novel aggregation layer (AG) (Sec. C.2.3) in order to i) aggregate the ground-truth from crowdvotes matrix, ii) compute the sensitivity and specificity of each annotator, and iii) jointly learn the classifier by back propagating the derivative of the loss function. We refer to this augmented architecture as *AggNet*.

C.2.3 Aggregation Layer (AG):

The straightforward method to aggregate labels annotated by users, is to employ majority voting (MV) [135]:

$$\mu = \begin{cases} 1 & \bar{y} \geq 0.5 \\ 0 & \bar{y} < 0.5 \end{cases} \quad (\text{C.2})$$

where $\bar{y} = \frac{1}{|P|} \sum_{j=1}^P y^j$ is the average label of P users. However, this strategy assumes all users to be on an equal level of trustworthiness.

In our framework, we integrate the method initially proposed by Raykar *et al.* [201], showing a good performance in many applications [196]. On top of our CNN architecture, we aggregate labels μ , estimate the sensitivity α^j and the specificity β^j for each annotator $j \in P$, and jointly learn the classifier. The method is solved using the well-known expectation-maximization (EM) algorithm and adapted to learn the softmax classifier as follows:

- **Initialization:** Using the crowdvotes matrix \mathbf{Y} , the aggregated labels μ_i initialized with majority voting, α^j and β^j are initially computed from μ_i .
- **E-Step:** Given the observation set \mathbf{D} and a current estimate of parameters $\psi := \{\alpha, \beta, \mu\}$, the conditional expectation is computed as

$$\mathbb{E}\{\ln Pr[\mathbf{D}, \mathbf{g}|\psi]\} = \sum_{i=1}^N \mu_i \ln p_i a_i + (1 - \mu_i) \ln(1 - p_i) b_i, \quad (\text{C.3})$$

where

$$a_i = \prod_{j=1}^P [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1-y_i^j},$$

$$b_i = \prod_{j=1}^P [\beta^j]^{1-y_i^j} [1 - \beta^j]^{y_i^j},$$

$$p_i = \sigma(z_i) = \frac{e^{z_i}}{\sum_{c=1}^C e^{z_{ic}}}, \text{ the output of softmax layer,}$$

$$z_i = \mathbf{w}^T \mathbf{x}_i, \text{ the output of FC layer,}$$

\mathbf{g} is the hidden variable (ground-truth),

and the expectation is with respect to $Pr[\mathbf{g}|\mathbf{D}, \psi]$.

Using Bayes' theorem, the aggregated labels μ_i can be computed as follows:

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i (1 - p_i)}. \quad (\text{C.4})$$

The loss function in our aggregation layer (AG) is defined as the Negative log-likelihood:

$$\mathcal{L}(\mu_i, p_i) = -\mathbb{E}\{\ln Pr[\mathbf{D}, \mathbf{g}|\psi]\}, \quad (\text{C.5})$$

- **M-Step:** Based on the observation set \mathbf{D} and the current estimate μ_i , the model parameters ψ can be computed by taking the derivative of \mathcal{L} with respect to each parameter and equate it to zero. The updates for α^j and β^j can be obtained as follows:

$$\alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}, \quad \beta^j = \frac{\sum_{i=1}^N (1 - \mu_i)(1 - y_i^j)}{\sum_{i=1}^N (1 - \mu_i)}, \quad (\text{C.6})$$

The softmax function is non-linear and the gradient with respect to parameter \mathbf{w} should be back propagated to the CNN layers [141]. For this purpose, we can employ the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial p_i} \frac{\partial p_i}{\partial z_i} \frac{\partial z_i}{\partial \mathbf{w}}, \quad (\text{C.7})$$

where

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{\mu_i - p_i}{p_i(1-p_i)}, \text{ the output of AG Layer,}$$

$$\frac{\partial p_i}{\partial z_i} = p_i(\delta_{ij} - p_j), \text{ the output of softmax layer}^6,$$

$$\frac{\partial z_i}{\partial w} = \mathbf{x}_i, \text{ the output of FC layer.}$$

Then, weights are updated using Stochastic Gradient Descent (SGD) [29].

It is notable that **E-Step** and **M-Step** are computed in forward and backward propagation respectively, which means one EM iteration per epoch. The refining process is stopped when the loss function output barely changes to avoid overfitting.

To this end, the aggregation method takes into account the sensitivity and specificity of each annotator to aggregate the labels. Furthermore, the algorithm is adapted to handle:

- **Trustworthiness:** Some crowdsourcing platforms provide the customer with a single accuracy score γ that each user achieved on a qualitative test for a specific task. It has been suggested by Raykar et al. [201] to model a prior distribution on sensitivity and specificity to trust some participants more than others. With only a single accuracy score as provided by our scenario, this is however not possible.
- **Missing Labels:** A common failure in crowdsourcing is that some users annotate a few samples only. If all these samples happen to fall within one single class, sensitivity or specificity remains *unknown*. Furthermore, the user who annotates few samples only might be equally or even more trusted than a user who annotates more samples.

Therefore, we reformulate α^j and β^j , without loss of generality, in such a way to augment the number of True Positives (TP) and True Negatives (TN) when the user has high confidence (i.e. accuracy score) as follows:

$$\alpha^j = \frac{\tau\gamma^j|N_p| + \sum_{i=1}^{N_p} \mu_i \delta_i^j}{\tau\gamma^j|N_p| + \sum_{i=1}^N \mu_i} = \frac{(1+\tau\gamma^j)TP}{(1+\tau\gamma^j)TP+FN},$$

$$\beta^j = \frac{\tau\gamma^j|N_n| + \sum_{i=1}^{N_n} (1-\mu_i)(1-\gamma_i^j)}{\tau\gamma^j|N_n| + \sum_{i=1}^N (1-\mu_i)} = \frac{(1+\tau\gamma^j)TN}{(1+\tau\gamma^j)TN+FP},$$
(C.8)

where γ^j is the accuracy score for a particular user and τ is the hyper-parameter that leverage the user's confidence. To avoid numerical issues, we set the sensitivity and specificity to 0.5 for *unknown* cases.

C.3 Experiments and Results

We have designed our experimental setup such that first, the proposed multi-scale CNN architecture is validated before evaluating the aggregated labels from the crowdvotes and validating the proposed augmented CNN (*AggNet*).

⁶The Kronecker delta, $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

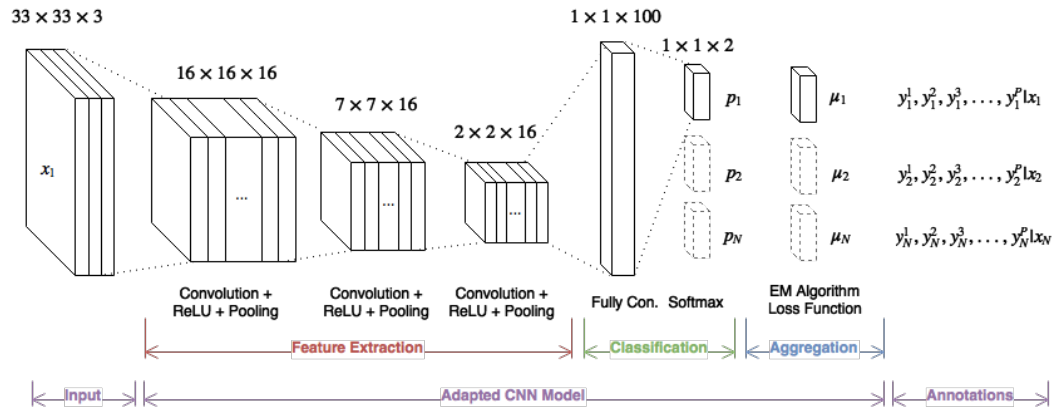


Figure C.2. *AggNet* architecture: The same CNN architecture is used for different scales, where p_i, μ_i, y_i^j represents the classifier output, the aggregated label, and the crowdvotes respectively.

Dataset. We have validated our proposed network on the publicly available MICCAI-AMIDA13 challenge dataset⁷. It contains annotated histology images of a total of 23 patients, who underwent invasive breast biopsy. During this medical examination, sections of suspicious breast tissue are collected and stained using hematoxylin and eosin (H&E). A histology RGB image of $2k \times 2k$ is then acquired with the Aperio ScanScope XT scanner at 40X magnification and with a spatial resolution of $0.25\mu\text{m}/\text{pixel}$. Then, a region of interest is identified and digitized to several high power field (HPF) images. The standard procedure in pathology is to count the mitotic figures in this area for the purpose of cancer grading (mitotic count score criteria). The annotation in the AMIDA13 dataset was done by two expert pathologists. Concordant annotations of both experts were taken as ground truth objects directly, whereas discordant cases were presented to two additional observers, such that the ground truth have been agreed upon by at least two experts. The reader is referred to [246] for more information about the dataset and its clinical/pathological background. In our experiment, we learn the proposed initial multi-scale model from 12 patients (311 HPF images), validate on 20% of the training set (60 HPF images) and test it on the whole testing data of AMIDA Challenge, including 11 patients (295 HPF images).

Implementation details. Each input RGB image is first pre-processed by staining appearance normalization [154]. Then, small patches of 33×33 are collected. Furthermore, to handle highly imbalanced data, patches showing positive classes are augmented with rotation and mirroring in such a way to leverages the ratio of positive to negative classes about (3:7). The multi-scale CNN is implemented using MATLAB and MatConvNet[242] and conducted on an Intel i7 machine with a GeForce GT 750M graphics card. Concerning the network parameters, the learning rate is set to 1×10^{-3} , momentum to 0.9, weight decay to 5×10^{-4} , and the batch size fixed to 200 samples. Note that some of these parameters are changed in the refining process, i.e. learning rate is set to 5×10^{-5} and the batch size is set to the whole crowdsourcing set. For the sake of reusability and to overcome the limitations of the crowdsourcing platform Crowdfunder, we have designed and implemented Annot8⁸, a Ruby-on-Rails based web-platform, allowing registered users to create datasets, upload images and labels, and categorize the labels with the help of a powerful tagging system. Collections of existing labels can be sent to and crowdsourced labels can be imported from Crowdfunder easily. Our web-platform also offers an online image processing frontend for on-demand patch extraction and

⁷AMIDA13: <http://amida13.isi.uu.nl/>

⁸Annot8: <http://vnmnavab14.informatik.tu-muenchen.de/>

Table C.1. DATASETS SPECIFICATIONS

	Proof of Concept	Use-Case
Model	8 Patients of Training set (318 HPF images)	The Entire Training set (371 HPF images)
Testing	3 Patients of Training set (22 HPF images)	The Entire Testing set (295 HPF images)
Crowdsourcing	Positive Candidates (550 Patches)	Positive Candidates (750 Patches)

computation of biomedical image filtering. On the participant side, each user was introduced briefly about the disease and the instructions of the actual task showing some good and bad examples as shown in Fig. C.3. Then, participants had to conduct a few test questions for quality control purposes. Without being made aware of the quiz mode, each annotator was presented with patches with known labels. Only then, he/she started to annotate five patches presented along with the filtered images. In order to ensure continuous quality control, a few randomly seeded test patches were still shown during the actual annotation job.

Evaluation metrics. We calculate different validation measures for comparison purpose, such as Recall = $\frac{TP}{TP+FN}$ and Precision = $\frac{TP}{TP+FP}$, where TP, FP and FN represent the true positives, false positives and false negatives respectively.

We further employ widely used statistical measures, such as F_1 -score = $\frac{2 \times TP}{FP+FN+2 \times TP}$, Receiver Operating Characteristics (ROC) and its Area Under Curve (AUC).

For measuring the improvement of multi-scale CNN over single scales, we compute the mean and standard variation of Relative Changes (RC) of different scales, i.e. $\mu_{RC} = \frac{1}{|Scales|} \sum_{Scales} \frac{MultiScale-xScale}{xScale}$.

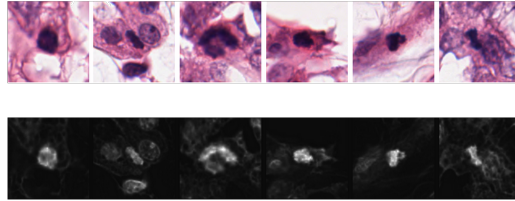
C.3.1 Proof-of-Concept Evaluation

The objective of this experimental setup is to analyze the functionality of the entire *AggNet* framework (c.f. Fig. C.1), specifically i) the accuracy of multi-scale CNN, ii) the performance of our novel aggregation layer when fed with noisy annotations, and iii) the influence of the augmented model on the detection quality of the multi-scale CNN. In order to perform quantitative analysis with respect to real ground-truth data, we decided to employ the AMIDA13 Challenge training dataset only as it comes with ground truth annotations (c.f. TABLE C.1). The setup is designed according to our overall framework pipeline depicted in Fig. C.1. First, a model is learned on 8 random patients of the AMIDA13 Challenge training dataset and tested on different 3 patients. Training and testing is performed employing our novel multi-scale CNN *AggNet*. Then, response maps of *AggNet* are thresholded at a lower operating point ensuring a large number of positive candidates, repatched and sent to CrowdFlower using our Annot8 web-platform.

In this experiment, we have crowdsourced around 550 patches, where each patch was annotated by 10 participants at least, resulting in more than 5500 labels stored in the crowdvotes matrix **Y**. We

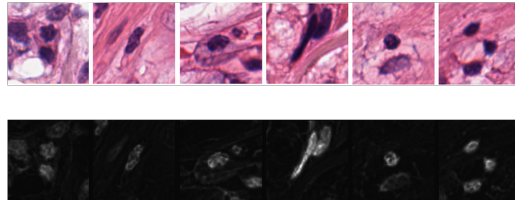
Tips and Examples

Mitosis:



The second row shows the corresponding so called "blueRatio" representation of the mitotic figures. Note how they have very bright spots!

Non-Mitosis



The second row shows the corresponding so called "blueRatio" representation of the non-mitotic figures. Note how they do not have such bright spots as the mitotic blue ratio representations!

Tips:

Mitotic figures usually look much more irregular than non-mitotic ones and appear to be very dark. In their blue ratio representation, they tend to have very bright spots. Watch out for very dark blobs that have a highly irregular (not really circle-like) shape and that have very bright spots in their blue ratio representation!

Figure C.3. Instructions and guidelines

have then evaluated results according to the different aspects related to the objectives defined for this specific experimental setup:

Multi-scale CNN

In order to measure the performance of the multi-scale CNN, we train the network on three different scales (0.33, 0.66 and 1) individually. For inference, we geometrically average the resulting FDMs from all scales to obtain the final positive responses. As alternative method to extract positive responses, we also threshold the FDM for each scale. For each remaining response, we check whether it is a TP, FP or FN detection. A positive response is considered a TP if its Euclidean distance to a ground-truth mitotic figure is less or equal than 30px. Multiple responses within the same radius around a mitotic figure are counted as a single TP. If there is no response for a mitotic figure within this radius, we count a FN detection. Any responses that are not inside any 30px region around a mitotic figure are counted as FP. It should be noted that a 30px radius ($7.5\mu\text{m}$) is used on the original scale, however, this is adjusted for different scales. Using these numbers, we calculate the Precision, Recall, and F_1 -score over all HPF slides at once and also per patient. TABLE C.2 shows the F_1 -score of 22 HPF images, the corresponding testing dataset, while the bar plot in Fig. C.5 displays the other metrics. It is obvious that the multi-scale CNN approach pushed the overall F_1 -score about $22.5\% \pm 6.8$, which validates our initial hypothesis of the proposed multi-scale CNN approach yielding a more robust classification due to detection consensus at various scales.

Aggregated Labels (AL)

To investigate the aggregated labels of the crowdvotes matrix \mathbf{Y} , we first run majority voting (MV) [135] and GLAD [273] methods on \mathbf{Y} without any quality control, referred to as MV-NoQ and GLAD-

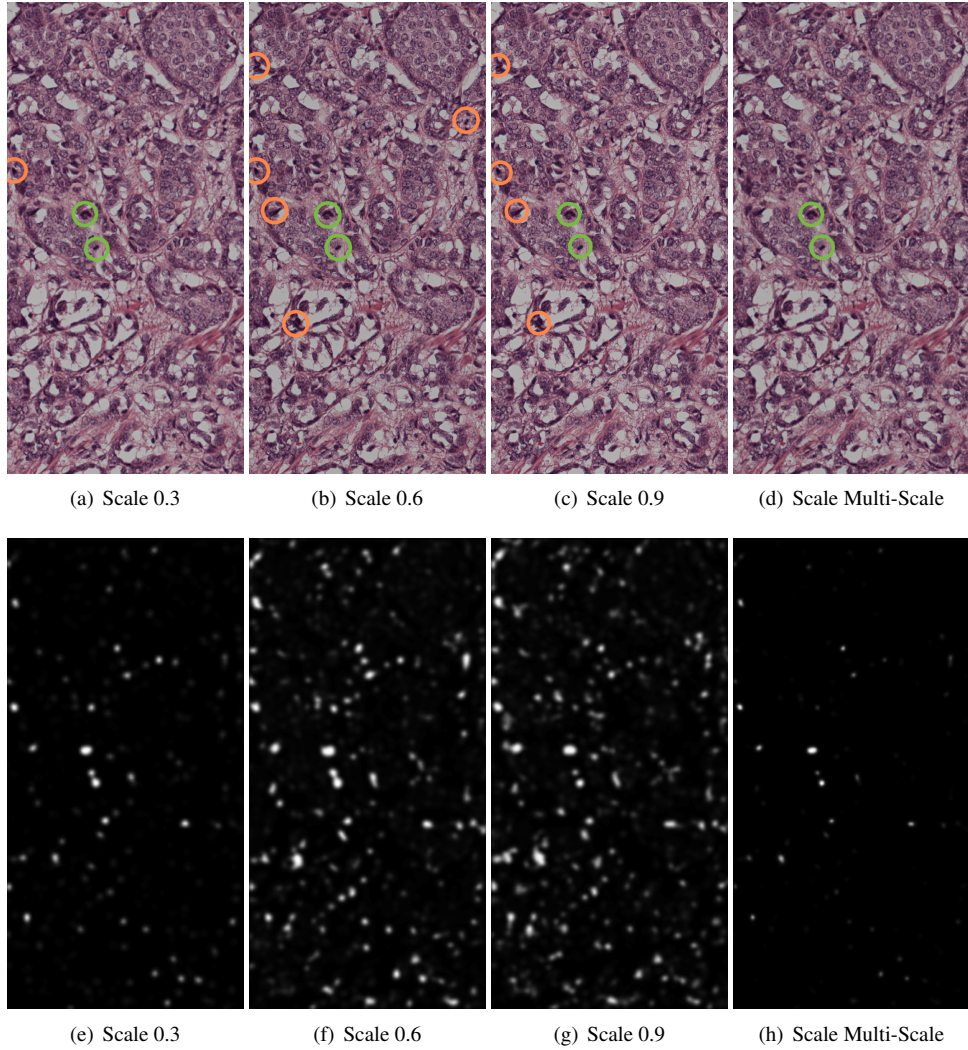


Figure C.4. First row shows results of one single image using multi-scale CNN, Green: the true positives, Orange: the false positives. Second row shows the corresponding final detection map (FDM) before thresholding. Best viewed in color.

NoQ respectively. Then, we use the existing 0.33-scale model from the previous multi-scale CNN experiment and augment it with our AG layer. Subsequently, we retrain it further for 100 epochs. We refer to this architecture as *AggNet*, and the aggregation results as AG-NoQ.

Further, to test the quality control, we filtered first the crowdvotes matrix \mathbf{Y} in order to keep only the annotations from the users who achieved more than 70% accuracy score in their qualitative test (in a quiz stage, each user had to annotate a few samples extracted from training data with known ground-truth). Then, we run the same aggregation methods, however, MV is replaced with Honeypot (HP) [142]. We refer to the aggregated labels using 70% quality control as HP-70Q, GLAD-70Q, and AG-70Q.

In addition, to figure out how the accuracy scores γ of the participant can influence the proposed aggregation method, we validate the hyper-parameter $\tau = [0.1, 0.2, \dots, 1]$ (ref. Sec. C.2.3) and run similar experiments, referred to as AG-NoQ- τ and AG-70Q- τ respectively.

Table C.2. F_1 -SCORES

	Patient 9	Patient 11	Patient 12	Overall
0.33-Scale	0.8000	0.5833	0.7778	0.6479
0.66-Scale	0.5000	0.5556	0.6957	0.5882
Orig.-Scale	0.5000	0.5490	0.6957	0.5854
Multi-Scale	0.8000	0.7368	0.7368	0.7419
Improvement	40%±36	39%±11	2.2%±6.4	22.5%±6.8

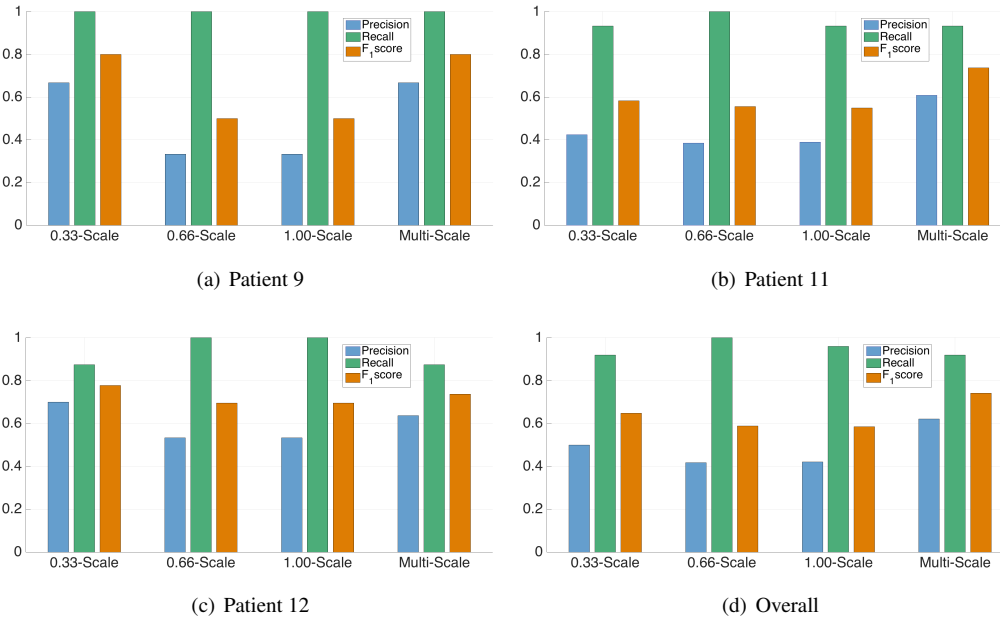


Figure C.5. Evaluation Metrics: Precision, Recall, and F_1 -score of Patients 9, 11 and 12.

To evaluate the aforementioned aggregated labels of crowdsourced patches (c.f. TABLE C.3), we compute the F_1 -score between the ground-truth and the aggregated labels of different methods as well as the proposed method (c.f. Fig. C.6(a)). Further, to give the reader more insight on the probabilistic scores of different aggregation methods, we plot the ROC curves of the significant methods as shown in Fig. C.7(a).

Unlike the weak agreement of the aggregated labels of both MV-NoQ and GLAD-NoQ with the ground-truth, both AG-NoQ and AG-NoQ- τ (-70Q as well) achieve an outperforming agreement at the first few epochs before getting decayed and saturated at still good agreement compared with the other aggregation methods as shown in Fig. C.6(a). MV-NoQ and GLAD-NoQ coincide in this setup due to the choice of our label threshold = 0.5. The ROC curve of the respective methods shows that GLAD is slightly superior to MV.

Interestingly, the aggregation methods without any quality control perform better than the ones with 70% quality control, which shows that the quality control strategy should be revised for such challenging data. Notably, AG-NoQ- τ as well as AG-70Q- τ , which incorporate the gamer accuracies

Table C.3. AGGREGATED LABELS

	Aggregation Method	Quality Control	Prior
MV-NoQ	MV	No	-
GLAD-NoQ	GLAD	No	-
AG-NoQ	Proposed	No	-
AG-NoQ- τ	Proposed	No	τ
HP-70Q	HP	70%	-
GLAD-70Q	GLAD	70%	-
AG-70Q	Proposed	70%	-
AG-70Q- τ	Proposed	70%	τ

Table C.4. AUGMENTED MODELS

	AM-GT	AM-MV	AM-GLAD	<i>AggNet</i>
F_1 -score	0.6250	0.6097	0.6097	0.6133
AUC	0.8433	0.8082	0.8082	0.8695

as a prior, underperform the other AG-models. Once again, this shows that the quality control strategy and thus the accuracy is a potential bottleneck.

Augmented Models (AM)

We further investigate how aggregated labels obtained from the previous experiment, influence the detection quality of our CNN compared to the ground-truth model. For this purpose, we compute several augmented models based on 0.33-scale of the initially trained ground truth model (GT). Besides *AggNet*, which we obtain by attaching our AG layer to GT, we compute three additional distinct models by retraining with aggregated labels from MV and GLAD as well as the real ground truth labels. We refer to the augmented models as AM-MV, AM-GLAD and AM-GT respectively. In fact, we retrain for 100 epochs, but pick the best performing model for each. Note that AM-GT is the 0.33-scale CNN model, however, its operating point is set to the same threshold that was used to select positive candidates for crowdsourcing. Once retrained, we utilize the models to perform mitosis detection on the corresponding testing set. Fig. C.7(b) and TABLE C.4 nicely outline how the proposed *AggNet* model almost performs as good as the augmented ground truth model AM-GT and easily outperforms both AM-MV and AM-GLAD (around 7.6% of AUC). Furthermore, some HPF images are visualized in TABLE C.5 to show how different augmented models perform on their optimal operating points. It is worth noticing also here how the proposed *AggNet* hits only the ground-truth in the perfect scenario outperforming even the AM-GT, or miss some ground-truth while having very few false positive in off scenario.

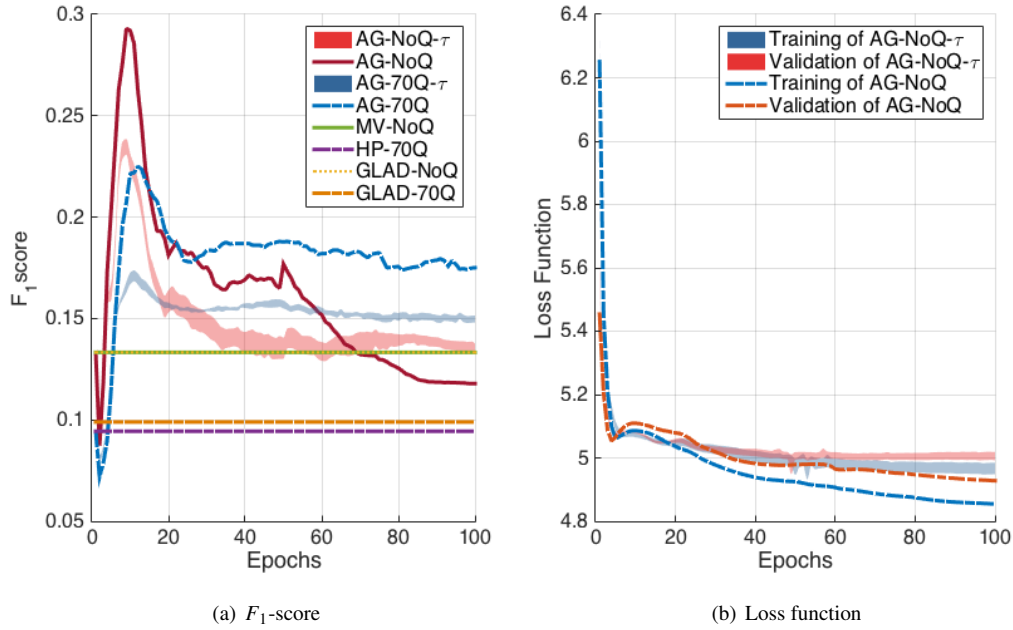


Figure C.6. (a) The aggregated labels of the crowdsourcing set are evaluated using the F_1 -score metric. (b) The loss function barely changes at 3-8 epochs before starting to overfit and the gap between the validation and training curves becomes significant. The shaded area depicts the change of τ .

C.3.2 Use Case Evaluation

This experiment aims at proving the overall impact of *AggNet* on a large standardized dataset (entire AMIDA13 Challenge training and testing datasets).

To first evaluate the performance of our proposed multi-scale CNN, we have participated in the AMIDA13 Challenge with our novel approach achieving 0.433 as overall F_1 -score. As reported by the challenge organizers, our method yields rank three of 15 participating methodologies.

Response maps resulting from the AMIDA13 Challenge testing dataset on 0.33-Scale have then been thresholded at an operating point of 0.99 (calculation based on the dataset) and repatched samples have been forwarded to CrowdFlower. Similarly to the previous experimental setup, crowdvotes \mathbf{Y} and confidences have been fed back to *AggNet* in order to augment the previously trained model. For performance evaluation, different augmented models have also participated in the AMIDA13 Challenge. TABLE C.7 shows the evaluation metrics of different models augmented with MV, GLAD and our robust aggregation layer (*AggNet*). The overall F_1 -score of *AggNet* easily outperforms the other augmented models of MV and GLAD, however, it falls slightly behind the previously trained model (0.33-Scale).

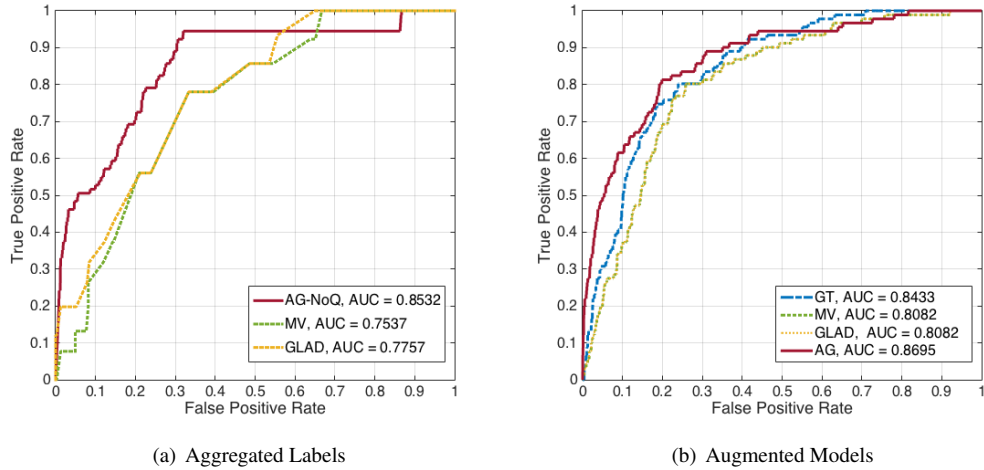


Figure C.7. ROC curves of the (a) aggregated labels using MV, GLAD and the proposed AG-NoQ, (b) the augmented models AM-GT, AM-MV, AM-GLAD and *AggNet* as well.

C.4 Discussion

Our results confirm that aggregation and deep learning from crowd annotations using the proposed *AggNet* is robust to "noisy" labels and positively influences the performance of our CNN in the refining phase.

Proper selection of operating point for crowdsourcing is quite challenging, for instance, it can be chosen based on the validation set, however, it might not be optimal for crowdsourcing where you need more positives. Therefore, it is recommended to plot the AUC, which provides the clinicians with more flexibility to run the model at the optimal operating point from their perspective. The augmented model *AggNet* has a gain of 7.6% in AUC at lower than the optimal operating point, however, this can not be computed for the use case experiment due to the limited number of submissions to the AMIDA Challenge.

However, the aggregation results from quality-filtered crowd annotations shed the light on the applied quality control. Surprisingly, they turned out to be worse than the aggregation from the complete, "noisy" set of crowd votes. This is clearly related to the high ambiguity and the high level of difficulty in detecting mitotic figures in general. Indeed, such quality tests need to be carefully planned and well designed in order to make sure they do not carry more "noise" than the actual crowdsourcing task, which we believe, the latter was the case in our experiments. However, this problem can also be related to the community of the crowd itself [245].

The low agreement among the crowd together with the small number of patches might be the reasons of the noticed decay in the aggregation F_1 -score curve, which leads to an overfitting after only few epochs (see Fig. C.6). Nevertheless, it is obvious that our robust aggregation layer can detect the novices (spammers) and weight their votes less. Therefore, our *AggNet* outperforms easily the other augmented models, where the spammers hurt the aggregations (c.f. TABLE C.7). Fig. C.8 shows the accuracy scores γ (based on the qualitative test) and the spammer scores S_p (based on the participant's

Table C.5. AGGREGATION AND DETECTION RESULTS

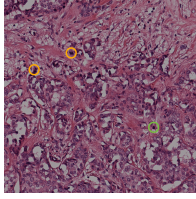
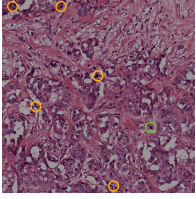
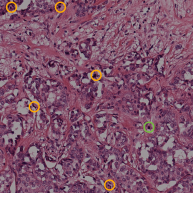
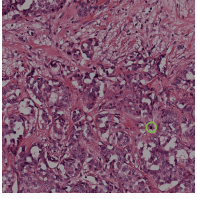
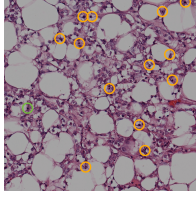
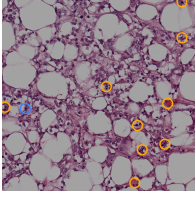
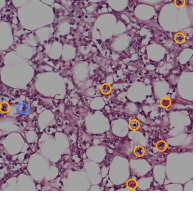
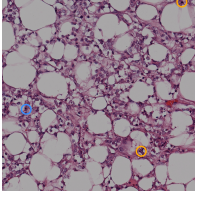
Augmented Models	AM-GT	AM-MV	AM-GLAD	The proposed <i>AggNet</i>
Perfect				
Off				

Table C.6. USE CASE RESULTS

	Precision	Recall	Overall F_1 -score
0.33-Scale	0.211	0.538	0.303
0.66-Scale	0.296	0.583	0.393
Orig.-Scale	0.172	0.400	0.241
Multi-Scale	0.441	0.424	0.433
Improvement	105%±54	-14%±18	44%±35

sensitivity α and specificity β , where $S_p = (\alpha + \beta - 1)^2$ [200]) of 100 participants in the crowdsourcing task.

During our research, the very natural question arose whether it may be possible to learn a model from crowdsourcing labels alone. For this purpose, we ran two additional crowdsourcing experiments. First, we utilized the crowdsourcing set (i.e. 5500 patches) and the binary classification crowdvotes to learn a model from scratch. Very soon, we realized that this model quickly overfitted due to the small number of training instances. Second, instead of publishing patches only, we asked the crowd to label full HPF images. In this case, however, due to insufficient settings within CrowdFlower platform, each participant labeled only few potential mitotic figures per image and left most of the challenging cases besides. This led to a large number of missing annotations and poor overall agreement among participants, which renders aggregation and training impossible. Still, these initial experiments gave us evidence that it is very difficult and maybe even impossible to entirely outsource the task of labeling mitotic figures in histology images to crowds. Instead, we decided to rather augment a small and narrow model learned from expert labels with wide but noisy crowd annotations to enhance variations encoded within the model.

Table C.7. USE CASE AUGMENTED MODELS

	0.33-Scale	AM-MV	AM-GLAD	<i>AggNet</i>
Precision	0.211	0.006	0.006	0.374
Recall	0.538	0.004	0.004	0.208
F_1 -score	0.303	0.004	0.005	0.267

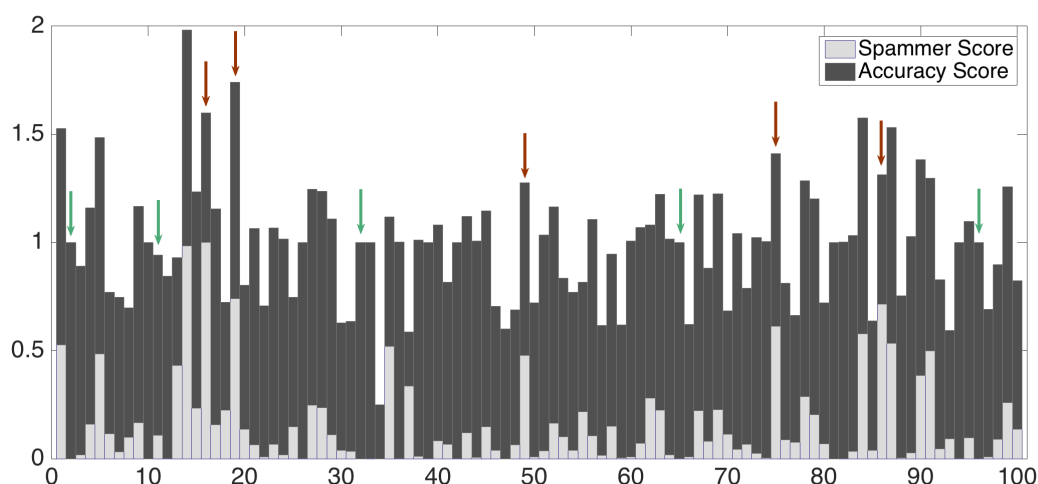


Figure C.8. Participants Analysis: accuracy and spammer scores of 100 participants. Arrows in green show some participants achieve high accuracy scores in the qualitative test, however, they are spammer. Arrows in red show very few participants who have good accuracy score as well as spammer score. Note that spammer score "0" means the participant is spammer.

Validating our methodology based on the smallest scale of our initial model is theoretically feasible. However, in future work, we want to conduct even more involved experiments including also the models of the other scales. This includes independent crowd sourcing rounds for each scale separately since retraining all the scales of the multi-scale model from the same crowd-sourced and aggregated labels hurts the concept of "redundancy & aggregation". Additionally, we want to consider multi-class classification which is, due to the binary nature of sensitivity and specificity, not directly possible, but can be performed in an one-vs-all fashion.

C.5 Conclusion

In this paper, we have introduced a novel concept for learning from crowds. Our new multi-scale CNN *AggNet* is designed to handle data aggregation directly as part of the learning process via an additional crowdsourcing layer. In our experimental study, we have further presented valuable insights into the functionality of deep CNN learning from crowd annotations and proven the impact of our novel aggregation scheme. To the best of knowledge, this is the first time that deep learning has been applied to generate a ground-truth labeling from non-expert crowd annotation in a biomedical context.

Although data aggregation is certainly necessary to learn from crowds, computational aggregation models have a limited impact, in particular if noisy crowd annotations are not significant, i.e. do not arise from ambiguous contexts. Besides clear guidelines, non-expert users need to be motivated to perform the task until the very end. Gamification is the ultimate solution here and we will focus future work on novel solutions on how to transform complex expert tasks in the biomedical domain into a game for non-expert users.

Acknowledgment

We would like to thank the reviewers for their constructive feedback and all the users who participated in this study. We are also very grateful to Dr. Mitko Veta for giving us the permission to use the AMIDA13 dataset in our research and supporting us during validation.

C.6 References

- [22] L. Aroyo and C. Welty. “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation”. In: *AI Magazine* 36.1 (2015) (cit. on p. 73).
- [26] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. “Robust Optimization for Deep Regression”. In: *Proc. IEEE Int. Conf. Computer Vision (ICCV)*. IEEE, 2015 (cit. on p. 73).
- [29] L. Bottou. “Large-scale machine learning with stochastic gradient descent”. In: *Proc. Computational Statistics*. 2010, pp. 177–186 (cit. on p. 77).
- [36] G. Carneiro, J. C. Nascimento, and A. Freitas. “The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods”. In: *IEEE Trans. Image Process.* 21.3 (2012), pp. 968–982 (cit. on p. 73).
- [43] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. “Mitosis detection in breast cancer histology images with deep neural networks”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. 2013, pp. 411–418 (cit. on p. 73).
- [65] D. Eigen, C. Puhrsch, and R. Fergus. “Depth map prediction from a single image using a multi-scale deep network”. In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2014 (cit. on p. 73).
- [69] E. Estellés-Arolas and F. González-Ladrón-De-Guevara. “Towards an Integrated Crowdsourcing Definition”. In: *J. Inf. Science* 38.2 (2012), pp. 189–200 (cit. on p. 72).
- [87] R. Girshick, J. Donahue, T. Darrell, and J. Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 2014 (cit. on p. 73).
- [117] J. Howe. “The rise of crowdsourcing”. In: *Wired Magazine* 14.6 (2006), pp. 1–4 (cit. on p. 72).
- [119] O. Inel, K. Khamkham, T. Cristea, et al. “CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data”. In: *Proc. Inter. Semantic Web Conf., Part II*. 2014, pp. 486–504 (cit. on p. 72).
- [129] F. Kleemann, G. Voß, and K. Rieder. “Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing”. In: *STI Studies* 4.1 (2008) (cit. on p. 72).

- [133] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2012, pp. 1097–1105 (cit. on pp. 4, 17, 72, 73).
- [135] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. Duin. “Limits on the majority vote accuracy in classifier fusion”. In: *Pattern Analysis & Applications* 6.1 (2003), pp. 22–31 (cit. on pp. 75, 80).
- [141] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. “Efficient BackProp”. In: *Neural Networks: Tricks of the Trade*. 2012, pp. 9–48 (cit. on p. 76).
- [147] F. Liu and L. Yang. “A Novel Cell Detection Method Using Deep Convolutional Neural Network and Maximum-Weight Independent Set”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. 2015, pp. 349–357 (cit. on p. 73).
- [148] J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 2015 (cit. on p. 73).
- [156] L. Maier-Hein, S. Mersmann, D. Kondermann, et al. “Crowdsourcing for reference correspondence generation in endoscopic images”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2014*. 2014, pp. 349–356 (cit. on pp. 72, 73).
- [177] V. Nair and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Proc. Int. Conf. Machine Learning (ICML)*. 2010, pp. 807–814 (cit. on p. 75).
- [196] N. Quoc Viet Hung, N. Tam, L. Tran, and K. Aberer. “An Evaluation of Aggregation Techniques in Crowdsourcing”. English. In: *Web Information Systems Engineering – WISE 2013*. 2013, pp. 1–15 (cit. on pp. 73, 75).
- [200] V. C. Raykar and S. Yu. “Ranking annotators for crowdsourced labeling tasks”. In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2011, pp. 1809–1817 (cit. on p. 86).
- [201] V. C. Raykar, S. Yu, L. H. Zhao, et al. “Learning from crowds”. In: *J. Machine Learn. Res.* 11 (2010), pp. 1297–1322 (cit. on pp. 73, 75, 77).
- [217] J. Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117 (cit. on p. 73).
- [219] V. S. Sheng, F. Provost, and P. G. Ipeirotis. “Get another label? improving data quality and data mining using multiple, noisy labelers”. In: *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. 2008, pp. 614–622 (cit. on p. 73).
- [242] A. Vedaldi and K. Lenc. “MatConvNet-Convolutional Neural Networks for MATLAB”. In: *Proc. ACM Int. Conf. Multimedia*. 2015 (cit. on p. 78).
- [245] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. “Community-based bayesian aggregation models for crowdsourcing”. In: *Proc. Int. Conf. World Wide Web*. ACM. 2014, pp. 155–164 (cit. on p. 85).
- [250] L. Von Ahn. “Games with a purpose”. In: *Computer* 39.6 (2006), pp. 92–94 (cit. on p. 72).
- [251] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. “recaptcha: Human-based character recognition via web security measures”. In: *Science* 321.5895 (2008), pp. 1465–1468 (cit. on p. 72).
- [273] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. “Whose vote should count more: Optimal integration of labels from labelers of unknown expertise”. In: *Advances in neural information processing systems*. 2009, pp. 2035–2043 (cit. on p. 80).

- [278] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang. “Beyond Classification: Structured Regression For Robust Cell Detection Using Convolutional Neural Network”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. 2015, pp. 358–365 (cit. on p. 73).
- [280] B. Yu, M. Willis, P. Sun, and J. Wang. “Crowdsourcing participatory evaluation of medical pictograms using Amazon Mechanical Turk”. In: *J. Med. Internet Res.* 15.6 (2013), e108 (cit. on p. 72).

Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research

Shadi Albarqouni[†], Stefan Matl[†], Maximilian Baust[†], Nassir Navab^{†*}, and Stefanie Demirci[†].

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

* Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Reprint Denied. The reprint of this publication was rejected on open-access platforms. The publication can be found at https://link.springer.com/chapter/10.1007/978-3-319-46976-8_28. The details are provided below.

Copyright Statement. ©2016 Springer and Deep Learning and Data Labeling for Medical Applications. LABELS 2016, DLMIA 2016. Lecture Notes in Computer Science, vol 10008. Springer, 2016, pp 269-277, Shadi Albarqouni, Stefan Matl, Maximilian Baust, Nassir Navab, and Stefanie Demirci, 'Playsourcing: A Novel Concept for Knowledge Creation in Biomedical Research'. Reprint denied.

Contributions. The author of this thesis was responsible for the main idea of gamifying the crowdsourcing task and transferring the extracted features to a game object, and for the validation and testing as well as for writing the manuscript. Stefan Matl, joint first author, was responsible for designing the game, validation and testing as well. Co-authors contributed to the revision of the manuscript.

Abstract. Being considered as a valid solution to the lack of ground truth data problem, crowdsourcing has recently gained a lot of attention within the biomedical domain. However, available concepts in life science domain require expert knowledge and thereby restrict the access to only very specific communities. In this paper, we go beyond state-of-the-art and present a novel concept for seamlessly embedding biomedical science into a common game canvas. Besides introducing the visual saliency concept, we thereby essentially eliminate the requirement for prior knowledge. We have further implemented a game to evaluate our novel concept in three different user studies.

Single-view X-ray depth recovery: toward a novel concept for image-guided interventions

Shadi Albarqouni[†], Ulrich Konrad[†], Lichao Wang[†], Nassir Navab^{†*}, and Stefanie Demirci[†].

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

* Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Reprint Denied. The reprint of this publication was rejected on open-access platforms. The publication can be found at <https://link.springer.com/article/10.1007/s11548-016-1360-0>. The details are provided below.

Copyright Statement. ©2016 Springer and International Journal of Computer Assisted Radiology and Surgery, Vol.11 (6), 2016, pp 873-880, Shadi Albarqouni, Ulrich Konrad, Lichao Wang, Nassir Navab, and Stefanie Demirci, 'Single-view X-ray depth recovery: toward a novel concept for image-guided interventions'.

Contributions. The author of this thesis was responsible for the main idea of estimating the depth from X-ray images, and the conceptualization as well as for writing the manuscript. Ulrich Konrad was responsible for the validation and testing as well. Co-authors contributed to the revision of the manuscript. Reprint denied.

Abstract. X-ray imaging is widely used for guiding minimally-invasive surgeries. Despite ongoing efforts in particular towards advanced visualization incorporating mixed reality concepts, correct depth perception from X-ray imaging is still hampered due to its projective nature. In this paper, we introduce a new concept for predicting depth information from single view X-ray images. Patient-specific training data for depth and corresponding X-ray attenuation information is constructed using readily available preoperative 3D image information. The corresponding depth model is learned employing a novel label consistent dictionary learning method incorporating atlas and spatial prior constraints to allow for efficient reconstruction performance. We have validated our algorithm on patient data acquired for different anatomy focus (abdomen and thorax). Of 100 image pairs per each of 6 experimental instances, 80 images have been used for training and 20 for testing. Depth estimation results have been compared to ground truth depth values. We have achieved around $4.40\% \pm 2.04$ and $11.47\% \pm 2.27$ mean squared error on abdomen and thorax datasets respectively, visual results of our proposed method are very promising. We have therefore presented a new concept for enhancing depth perception for image guided interventions.

X-ray In-Depth Decomposition: Revealing The Latent Structures

Shadi Albarqouni[†], Javad Fotouhi^{*}, and Nassir Navab^{†*}

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

^{*} Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Reprint Denied. The reprint of this publication was rejected on open-access platforms. The publication can be found at https://link.springer.com/chapter/10.1007/978-3-319-66179-7_51. The details are provided below.

Copyright Statement. ©2017 Springer and Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017, Lecture Notes in Computer Science Volume 10435, 2017, pp 444-452, Shadi Albarqouni, Javad Fotouhi, and Nassir Navab, 'X-ray In-Depth Decomposition: Revealing The Latent Structures'. Reprint denied.

Contributions. The author of this thesis was responsible for the main idea, the conceptualization, validation and testing as well as for writing the manuscript. Co-authors contributed to the revision of the manuscript.

Abstract. X-ray is the most readily available imaging modality and has a broad range of applications that spans from diagnosis to intra-operative guidance in cardiac, orthopedics, and trauma procedures. Proper interpretation of the hidden and obscured anatomy in X-ray images remains a challenge and often requires high radiation dose and imaging from several perspectives. In this work, we aim at decomposing the conventional X-ray image into d X-ray components of independent, non-overlapped, clipped sub-volume, that separate rigid structures into distinct layers, leaving all deformable organs in one layer, such that the sum resembles the original input. Our proposed model is validated on 6 clinical datasets (~7200 X-ray images) in addition to 615 real chest X-ray images. Despite the challenging aspects of modeling such a highly ill-posed problem, exciting and encouraging results are obtained paving the path for further contributions in this direction.

Abstracts of Publications not Discussed in this Thesis

Structure-Preserved Color Normalization for Histological Images

Abhishek Vahadane^{*†}, Tingying Peng[†], Shadi Albarqouni[†], Maximilian Baust[†], Katja Steiger[◊], Anna Melissa Schlitter[◊], Amit Sethi^{*}, Irene Esposito[‡], and Nassir Navab^{†*}.

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

^{*} Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

^{*} Indian Institute of Technology Guwahati, India

[◊] Institute of Pathology, Technical University of Munich, Germany

[‡] Institute of Pathology, Medical University of Innsbruck, Austria

Copyright Statement. ©2015 IEEE. Reprinted, with permission from Abhishek Vahadane, Tingying Peng, Shadi Albarqouni, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Amit Sethi, Irene Esposito, and Nassir Navab, 'Structure-Preserved Color Normalization for Histological Images', In Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, pp. 1012-1015. IEEE, 2015.

Contributions. The author of this thesis was responsible for the implementation of dictionary learning in this paper.

Abstract. Automated image processing and quantification are increasingly gaining attention in the field of digital pathology. However, a common problem that encumbers computerized analysis is the color variation in histology, due to the use of different microscopes/scanners, or inconsistencies in tissue preparation. In this paper, we present a novel color normalization technique to bring a histological image (source image) into a different color appearance of a second image (target image), which therefore standardizes the color representation of both images. In particular, by incorporating biological stain-sparse regularized stain separation, our color normalization technique preserves the structural information of the source image after color normalization, which is very important for subsequent image analysis. Both qualitative and quantitative validation demonstrates the superior performance of our stain separation and color normalization techniques.

Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images

Abhishek Vahadane^{*†}, Tingying Peng[†], Amit Sethi^{*}, Shadi Albarqouni[†], Lichao Wang[†], Maximilian Baust[†], Katja Steiger[◊], Anna Melissa Schlitter[◊], Irene Esposito[‡], and Nassir Navab^{†*}.

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

^{*} Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

^{*} Indian Institute of Technology Guwahati, India

[◊] Institute of Pathology, Technical University of Munich, Germany

[‡] Institute of Pathology, Medical University of Innsbruck, Austria

Copyright Statement. ©2016 IEEE. Reprinted, with permission from Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab, 'Structure-preserving color normalization and sparse stain separation for histological images', IEEE transactions on medical imaging, 35(8), pp.1962-1971.

Contributions. The author of this thesis was responsible for the implementation of dictionary learning in this paper.

Abstract. Staining and scanning of tissue samples for microscopic examination is fraught with undesirable color variations arising from differences in raw materials and manufacturing techniques of stain vendors, staining protocols of labs, and color responses of digital scanners. When comparing tissue samples, color normalization and stain separation of the tissue images can be helpful for both pathologists and software. Techniques that are used for natural images fail to utilize structural properties of stained tissue samples and produce undesirable color distortions. The stain concentration cannot be negative. Tissue samples are stained with only a few stains and most tissue regions are characterized by at most one effective stain. We model these physical phenomena that define the tissue structure by first decomposing images in an unsupervised manner into stain density maps that are sparse and non-negative. For a given image, we combine its stain density maps with stain color basis of a pathologist-preferred target image, thus altering only its color while preserving its structure described by the maps. Stain density correlation with ground truth and preference by pathologists were higher for images normalized using our method when compared to other alternatives. We also propose a computationally faster extension of this technique for large whole-slide images that selects an appropriate patch sample instead of using the entire image to compute the stain color basis.

CathNets: Detection and Single-View Depth Prediction of Catheter Electrodes

Christoph Baur[†], Shadi Albarqouni[†], Stefanie Demirci[†], Nassir Navab^{†*}, and Pascal Fallavollita[†]

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

* Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Copyright Statement. ©2016. MIAR. Reprinted, with permission from Christoph Baur, Shadi Albarqouni, Stefanie Demirci, Nassir Navab, and Pascal Fallavollita, 'CathNets: Detection and Single-View Depth Prediction of Catheter Electrodes', In International Conference on Medical Imaging and Virtual Reality, pp. 38-49. Springer International Publishing, 2016.

Contributions. The author of this thesis was responsible for the main idea. In addition to writing the manuscript.

Abstract. The recent success of convolutional neural networks in many computer vision tasks suggests that their application could also be beneficial for vision tasks in cardiac electrophysiology procedures which are commonly carried out under guidance of C-arm fluoroscopy. Many efforts for catheter detection and reconstruction have been made, but especially realtime and robust detection of catheters in X-ray images is still not entirely solved. We propose two novel methods for i) fully automatic electrophysiology catheter electrode detection in X-ray images and ii) depth estimation of such electrodes based on convolutional neural networks. For i), experiments on a total of 1650 X-ray images from 24 sequences yielded a detection rate $> 99\%$. Our experiments on ii) depth prediction using 20 images with depth information available revealed that we are able to estimate the depth of catheter tips in the lateral view with a remarkable mean error of $6.08 \pm 4.66\text{mm}$.

X-ray PoseNet: 6 DoF Pose Estimation for Mobile X-ray Devices

Mai Bui^{†‡}, Shadi Albarqouni[†], Michael Schrapp[‡], Nassir Navab^{†*}, and Slobodan Ilic[‡]

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

^{*} Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

[‡] Siemens AG, Munich, Germany

Copyright Statement. ©2017 IEEE. Reprinted, with permission from Mai Bui, Shadi Albarqouni, Michael Schrapp, Nassir Navab, and Slobodan Ilic, 'X-ray PoseNet: 6 DoF Pose Estimation for Mobile X-ray Devices', In IEEE WACV, 2017.

Contributions. The author of this thesis and Mai Bui are equally contributed to this work.

Abstract. Precise reconstruction of 3D volumes from X-ray projections requires precisely pre-calibrated systems where accurate knowledge of system geometric parameters is known ahead. However, when dealing with mobile X-ray devices such calibration parameters are unknown. Joint estimation of system calibration parameters and 3d reconstruction is heavily unconstrained problem, especially when the projections are arbitrary. In industrial applications, that we target here, nominal CAD models of object to be reconstructed are usually available. We rely on this prior information and employ Deep Learning to learn the mapping between simulated X-ray projections and its pose. Moreover, we introduce the reconstruction loss in addition to the pose loss to further improve the reconstruction quality. Finally, we demonstrate the generalization capabilities of our method in case where poses can be learned on instances of the objects belonging to the same class, allowing pose estimation of unseen objects from the same category, thus eliminating the need for the actual CAD model. We performed exhaustive evaluation demonstrating the quality of our results on both synthetic and real data.

Semi-supervised Learning for Fully Convolutional Networks

Christoph Baur[†], Shadi Albarqouni[†], and Nassir Navab^{†*}

[†] Computer Aided Medical Procedures (CAMP),
Technische Universität München, Germany

* Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA

Copyright Statement. ©2017 Springer and Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017, Lecture Notes in Computer Science Volume 10435, 2017, pp 311-319, Christoph Baur, Shadi Albarqouni, and Nassir Navab, 'Semi-supervised Learning for Fully Convolutional Networks'

Contributions. The author of this thesis and Christoph Baur are equally contributed to this work.

Abstract. Deep learning usually requires large amounts of labeled training data, but annotating data is costly and tedious. The framework of semi-supervised learning provides the means to use both labeled data and arbitrary amounts of unlabeled data for training. Recently, semi-supervised deep learning has been intensively studied for standard CNN architectures. However, Fully Convolutional Networks (FCNs) set the state-of-the-art for many image segmentation tasks. To the best of our knowledge, there is no existing semi-supervised learning method for such FCNs yet. We lift the concept of auxiliary manifold embedding for semi-supervised learning to FCNs with the help of Random Feature Embedding. In our experiments on the challenging task of MS Lesion Segmentation, we leverage the proposed framework for the purpose of domain adaptation and report substantial improvements over the baseline model.

