Deep Representation Learning from Imbalanced Medical Imaging

Dissertation

zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachgebiet Internet-Technologien und -Systeme
des Hasso-Plattner-Institute

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

vorgelegt von Master of Science

Mina Rezaei

Potsdam, June. 2019

Dissertation Reviewers:

Prof. Dr. Christoph Meinel, Hasso Plattner Institute
Prof. Dr. Nassir Navab, Technical University Munich
Prof. Dr. Heinz Handels, University of Lübeck

Examination Committee:

Prof. Dr. Felix Naumann, (Chairman)
Prof. Dr. Christoph Lippert,
Prof. Dr. Tilmann Rabl,
...and the reviewers

Submission: 06.06.2019
Disputation: 16.12.2019

Published online at the
Institutional Repository of the University of Potsdam:
https://doi.org/10.25932/publishup-44275
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-442759

برای دخترم، ویونا دیانا سوفیا

برای پسرم، پارسا کسرا آیین

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified. Notions taken over directly or indirectly from other sources have been identified as such. All data and findings in the work have not been falsified or embellished. This thesis has not previously been presented in identical or similar form to any other German or foreign examination board.

The thesis work was conducted from *November. 2015* to *June. 2019* under the supervision of *Professor Meinel* at *Hasso Plattner Institute*.

Mina Rezaei
Potsdam, Germany

Abstract

Medical imaging plays an important role in disease diagnosis, treatment planning, and clinical monitoring. One of the major challenges in medical image analysis is imbalanced training data, in which the class of interest is much rarer than the other classes. Canonical machine learning algorithms suppose that the number of samples from different classes in the training dataset is roughly similar or balance. Training a machine learning model on an imbalanced dataset can introduce unique challenges to the learning problem.

A model learned from imbalanced training data is biased towards the high-frequency samples. The predicted results of such networks have low sensitivity and high precision. In medical applications, the cost of misclassification of the minority class could be more than the cost of misclassification of the majority class. For example, the risk of not detecting a tumor could be much higher than referring to a healthy subject to a doctor. The current Ph.D. thesis introduces several deep learning-based approaches for handling class imbalanced problems for learning multi-task such as disease classification and semantic segmentation.

At the data-level, the objective is to balance the data distribution through re-sampling the data space: we propose novel approaches to correct internal bias towards fewer frequency samples. These approaches include patient-wise batch sampling, complimentary labels, supervised and unsupervised minority oversampling using generative adversarial networks for all.

On the other hand, at algorithm-level, we modify the learning algorithm to alleviate the bias towards majority classes. In this regard, we propose different generative adversarial networks for cost-sensitive learning, ensemble learning, and mutual learning to deal with highly imbalanced imaging data.

We show evidence that the proposed approaches are applicable to different types of medical images of varied sizes on different applications of routine clinical tasks, such as disease classification and semantic segmentation. Our various implemented algorithms have shown outstanding results on different medical imaging challenges.

Zusammenfassung

Medizinische Bildanalyse spielt eine wichtige Rolle bei der Diagnose von Krankheiten, der Behandlungsplanung, und der klinischen Überwachung. Eines der großen Probleme in der medizinischen Bildanalyse ist das Vorhandensein von nicht ausbalancierten Trainingsdaten, bei denen die Anzahl der Datenpunkte der Zielklasse in der Unterzahl ist. Die Aussagen eines Modells, welches auf einem unbalancierten Datensatz trainiert wurde, tendieren dazu Datenpunkte in die Klasse mit der Mehrzahl an Trainingsdaten einzuordnen. Die Aussagen eines solchen Modells haben eine geringe Sensitivität aber hohe Genauigkeit. Im medizinischen Anwendungsbereich kann die Einordnung eines Datenpunktes in eine falsche Klasse Schwerwiegende Ergebnisse mit sich bringen. In die Nichterkennung eines Tumors Beispielsweise brigt ein viel höheres Risiko für einen Patienten, als wenn ein gesunder Patient zum Artz geschickt wird.

Das Problem des Lernens unter Nutzung von nicht ausbalancierten Trainingsdaten wird erst seit Kurzem bei der Klassifizierung von Krankheiten, der Entdeckung von Tumoren und beider Segmentierung von Tumoren untersucht. In der Literatur wird hier zwischen zwei verschiedenen Ansätzen unterschieden: datenbasierte und algorithmische Ansätze. Die vorliegende Arbeit behandelt das Lernen unter Nutzung von unbalancierten medizinischen Bilddatensätzen mittels datenbasierter und algorithmischer Ansätze.

Bei den datenbasierten Ansützen ist es unser Ziel, die Datenverteilung durch gezieltes Nutzen der vorliegenden Datenbasis auszubalancieren. Dazu schlagen wir neuartige Ansätze vor, um eine ausgeglichene Einordnung der Daten aus seltenen Klassen vornehmen zu können. Diese Ansätze sind unter anderem *synthesize minority class sampling*, *patient-wise batch normalization*, und die Erstellung von komplementären Labels unter Nutzung von generative adversarial networks. Auf der Seite der algorithmischen Ansätze verändern wir den Trainingsalgorithmus, um die Tendenz in Richtung der Klasse mit der Mehrzahl an Trainingsdaten zu verringern. Dafür schlagen wir verschiedene Algorithmen im Bereich des kostenintensive Lernens, Ensemble-Lernens und des gemeinsamen Lernens vor, um mit stark unbalancierten Trainingsdaten umgehen zu können.

Wir zeigen, dass unsere vorgeschlagenen Ansätze für verschiedenste Typen von medizinischen Bildern, mit variierender Größe, auf verschiedene Anwendungen im klinischen Alltag, z. B. Krankheitsklassifizierung, oder semantische Segmentierung, anwendbar sind. Weiterhin haben unsere Algorithmen hervorragende Ergebnisse bei unterschiedlichen Wettbewerben zur medizinischen Bildanalyse gezeigt.

Acknowledgements

This thesis summarizes the scientific contributions of my graduate journey at HPI. However, alongside the research, this thriving environment has enabled more personal and professional growth than I could have ever imagined. With these next few paragraphs, I want to express my gratitude to the many people who have made this an exceptional experience that I will forever cherish. These few sentences do not do justice to the impact, they have had on me, but I hope to pay forward the technical depth and warm kindness I have been taught.

I am deeply grateful to Professor Christoph Meinel, an incredible mentor in all aspects of my graduate development. He has given me a lot of support and invaluable advice, making sure we learn to be excellent researchers while being good engineer. Professor Meinel, I am greatly indebted to you for sharing some of your valuable time, and level professionalism never ceases to impress. You supported me a lot not only for the research work and also in my daily life in Germany in the past several years. Danke Schön.

I want to thank Dr. Haojin Yang for his thorough attention to my work and thesis and his technical support. He always takes the time to listen and give scholarly and professional guidance, and his advice has helped me think more clearly about a raw idea, implementation, and writing a good paper. xie xie.

I want to thank Professor Hiro Yoshida for his attention to my work and his ongoing support. He takes the time to listen to my idea that I pitched and give professional guidance regarding clinical perspective through Skype meeting for several hours during last year. Arigato gozaimas.

I am fortunate to have had HPI-Research School scholarship, connecting with amazing people from a different research field in computer science. During almost four years, Wednesday afternoon weekly meeting, I learn how the other shapes, build, develop their Ph.D. project, Robert Kovacs, Martin Krejca, Sona Ghahremani, Thijs Roumen, Cheng Wang, Mirela Alistar, Alexandra Ion, David Mottin, Aragats Amirkhanyan, Julian Risch, Francesco Quinzan, and... I want to thank Professor Andreas Polze, Professor Robert Hirschfeld, Professor Felix Neumann, and Professor Patrick Baudish for their commitment to my scholarly

and personal development. I truly felt that they were invested in my development and well-being, which is rare and valuable to have.

I want to thank Professor Christoph Lippert for his attention to my work. Thank you for involving me for writing BMBF project. It was a great experience. I appreciate your time to listen to my ideas, plans, and decisions that I pitched to you and give professional guidance.

I want to thank Mrs. Sabine Wagner, Mrs. Michaela Schmitz, and Mrs. Daniela Roick that helped and assisted me for documents.

Most of the work presented in this thesis was done in close collaboration with wonderful colleagues, office mates, and friends, Haojin Yang, Goncalo Mordido, Christian Bartz, and Cheng Wang, I learned and inspired from you a lot.

I want to thank Tatiana and Hanadi for unlimited support, encouragement, and truly friend. I would like to thanks the colleagues who made joyful time in HPI including Nuhad, Anja, Julia, Pejman, Ali, Raad, Ting, Harry, Goncalo, Ihsan, Kennedy, and Johannes I believe these are truly rare connections.

I want to thank the student co-workers, who have worked for supporting this thesis. This includes Konstantin Harmuth, Dr. Sharon Nameth.

My parents Nahid Bagheri and Mehdi Rezaei, and my sisters, Mozhgan and Mahsa, have offered tireless support and encouragement. I know that their efforts and sacrifices lead to an opportunity for me to study and thrive at HPI. They raised me to be passionate and dedicated to fair and good work, and I will be eternally grateful for everything they have done for me. I am very happy to have enthusiastic conversations with my extended family, especially my inspirational grandparents. I am also amazed by the ongoing support offered by my husbands family Ali, Elahe, Moha, and Sara and will always be thankful to them.

Yaser (Abouzar), my husband, is my love, my partner in everything I do. Hes carried me through my Ph.D., from the small things - the infinitely many decisions to the big challenges. He spend almost weekend in Deutsche bahn from Munich - Berlin - Munich to meet each other for few days. However, we were geographically separated for 677 km, 187 weeks, 1309 days, 31416 hours, we have always remained close to each other. He is my pillar of support in all aspects of life, my source of clinical clarification and scientific wisdom. I always appreciate the ability to have rich scientific conversations of all sorts. I am grateful for his

love, understanding, and unconditional support, and for helping me improve into the scientist, I am today. His kindness, curiosity, and excitement will always be my inspiration and drive in life.

Thank you all for helping to shape me at such an important time in my life.

Contents

List of Figures

List of Tables

Chapter 1

Learning Representation from Imbalanced Medical Imaging

1.1 Introduction

Medical imaging such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound, and microscopic light imaging creates a visual representation of human body which provides essential information for disease diagnosis, treatment planning, and clinical monitoring. The main goal of research on medical image analysis is to extract clinic relevant information or knowledge from medical images. Automated image segmentation is an important and challenging clinical routine task which identifies the boundaries of the region of interest (ROI) such as body organ or abnormal tissues in images. The segmentation result provides critical knowledge for shape analysis, detecting changes in the volume, planning for radiation therapy, etc. Hence, manual annotation is very time consuming and subjective, an accurate and reliable automated segmentation method is valuable for both clinical and research purposes.

Machine learning methods for automated medical image segmentation can be divided into two main categories [2]. The first category is generative models that learn the joint distribution between features and corresponding labels. The other group is a discriminative model that learn conditional probability between features and targets.

1. LEARNING REPRESENTATION FROM IMBALANCED MEDICAL IMAGING

Generative probabilistic approaches build the model based on prior domain knowledge about the appearance and spatial distribution of the different tissue types. Traditionally, generative probabilistic models have been popular where simple conditionally independent Gaussian models [6] or Bayesian learning [7] are used for tissue appearance.

On the contrary, discriminative probabilistic models, also called conditional models, directly learn the relationship between the local features of images and segmentation labels without any domain knowledge. Traditional discriminative approaches such as SVMs [8], random forests [9] have been used in medical image segmentation.

Recently, generative adversarial networks (GAN) [10] have been successfully used for high-fidelity natural image synthesis, improving learned image compression, data augmentation, and more. Figure 1.1 shows the number of papers published based on GAN since 2014. The main idea behind conventional GANs is to train two neural networks: the generator, which learns how to synthesize data (such as an image), and the discriminator, which learns how to distinguish real data from the ones synthesized by the generator. For natural image synthesis, state-of-the-art results are achieved by conditional GANs [11] that has the advantage of being able to provide better representations for multi-modal data generation by control on the modes of the data being generated. This makes cGANs suitable for image semantic segmentation tasks, where we condition an input image and generate a corresponding output image. The conditional GAN can provide promising results for medical image segmentation since it doesn't need large training samples.

A common problem in clinical application of machine learning or deep learning based classifier is that some classes have a significantly higher number of examples in the training set than other classes. This difference is referred to as class imbalance. Even in the task of medical image semantic segmentation, the number of pixels or voxels belong to healthy class majority, and the lesion or non-healthy pixels or voxels are minor. A deep learning model trained on imbalanced data biased towards a majority class, which is healthy. The predicted results of such networks have low sensitivity where sensitivity shows the ability of a test to

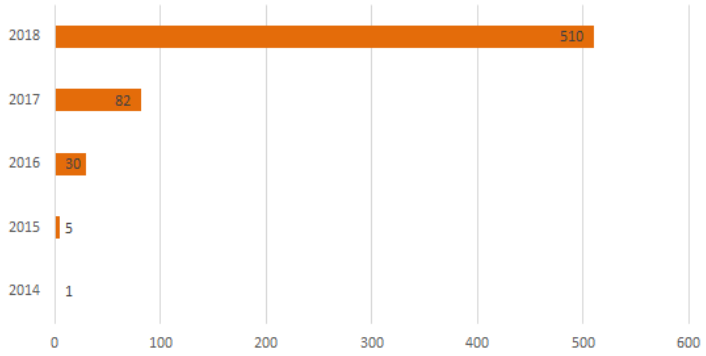


Figure 1.1: Number of published papers in generative adversarial net since 2014.

predict non-healthy classes correctly. In medical applications, the cost of misclassification of the minority class could be more than the cost of misclassification of the majority class.

Methods for mitigating class imbalance can be divided into two main categories:

- Data-level approaches that modify the class distribution through re-sampling the data space, by including SMOTE (Synthetic Minority Over-sampling Technique) of the positive class or by under-sampling of the negative class.
- Algorithm-level methods modify the learning algorithm to alleviate the bias towards the majority class. Examples are ensemble learning, cost-sensitive approaches, and one class learning.

Preview of Contributions

In this thesis, we present novel approaches to mitigate the class imbalance issue in both data-level and algorithm-level. We approach patient-wise batch normalization, and synthetic minority oversampling to skew internal bias from majority class using generative adversarial networks for both. We introduce different novel generative adversarial frameworks for cost-sensitive learning, ensemble learning, and mutual learning to deal with highly imbalanced imaging data. We show evidence that the proposed approaches are applicable to different types of medical images of varied sizes on different applications of routine clinical tasks, such

as diseases classification and semantic segmentation. Our various implemented algorithms have shown outstanding results on some recent medical imaging challenges.

1.2 Medical Image Analysis

Medical Imaging is the technique and process of creating a visual representation of the function of the interior bodies such as organs or tissues for clinical analysis and medical interventions. Medical imaging is a crucial step to reveal the internal structure of tissues and bones for diagnosing diseases and treatment planning. In clinical discipline, the medical practitioner responsible for interpreting and acquiring the image is radiologist, dermatologist, or medical expert.

Recently, Litjens et al. [12] review and summarize 300 papers with the major of deep learning and medical image analysis until June 2017. We added 300 high cited papers considering the same major of machine learning application on medical imaging published from 2017 till June 2019 in MICCAI, ISBI, SPIE, IPMI, MIDL conferences and workshops as well as IEEE Transaction Medical Imaging and Medical Imaging journals to survey list.

Figures 1.2,1.3, and 1.4 represent the statics in details of publication year, image modality, and application area. Medical image segmentation using machine learning and deep learning is the most popular research topic as shown in Figure 1.4, which classification and detection are second and third, respectively.

Image segmentation is a fundamental problem in medical image computing, which attempts to identify the exact boundaries of regions such as anatomical organs or abnormal tissues (e.g., lesion). Manual segmentation is time-consuming and tedious. Moreover, manual segmentation is subject to variation, between both observers and the same observer. An accurate automatic medical image segmentation is valuable in the clinical facility. Segmentation is the most common subject of papers applying deep learning to medical imaging, as shown in Figure 1.4.

Image segmentation is often a crucial first step in computer-aided detection pipelines. The segmentation of organs and other substructures in medical images

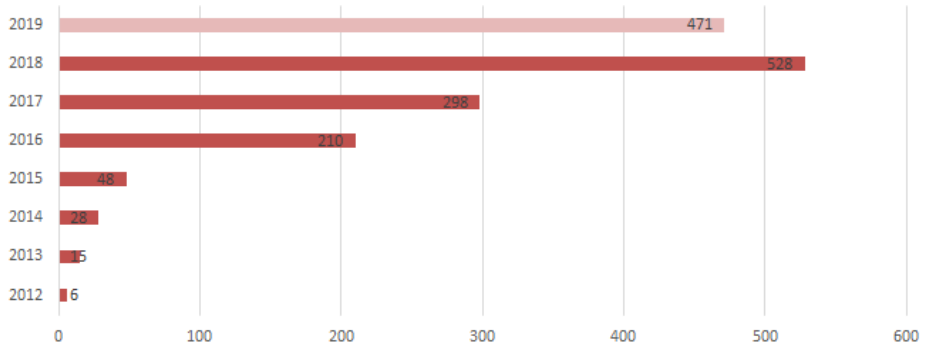


Figure 1.2: Number of medical image analysis papers using deep learning approaches published till June 2019.

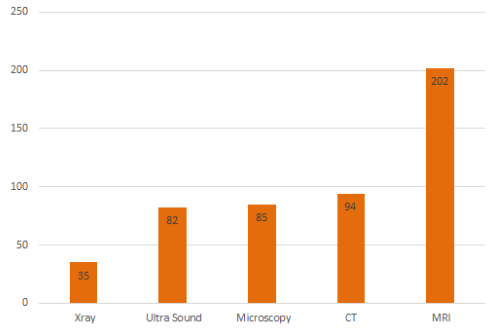


Figure 1.3: Categorization of deep learning research papers according to medical image modality

allows quantitative analysis of clinical parameters related to volume and shape, as, for example, in cardiac or brain study which as shown by Figure 1.5.

1.2.1 Image Acquisition

Depending on the patient and emergency the medical expert decides for one or more different type of medical imaging modalities such as ultrasound, Magnetic Resonance (MR), and Computed Tomography (CT) facilitate visualization of the human body, a central component of clinical practice. For example, MRI and CT scans are part of the standard care in diagnosing and evaluating most soft tissues. We provide more detail about the modalities that we've used in this thesis.

Computed Tomography (CT)

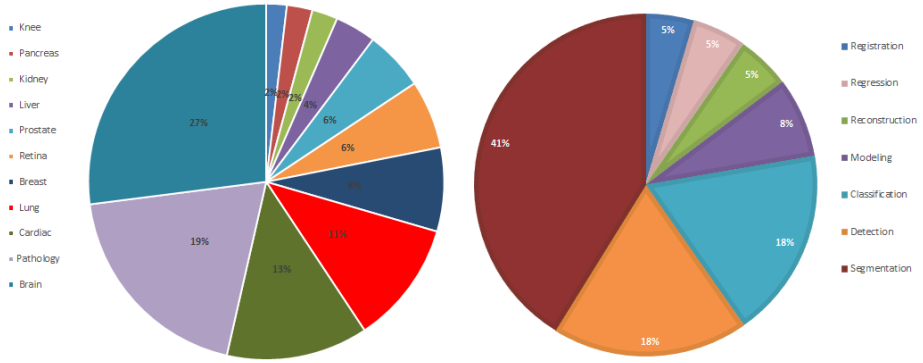


Figure 1.4: The distribution of research papers in the different body structure depicted in left side and the percentage of research among various tasks shown in the right side

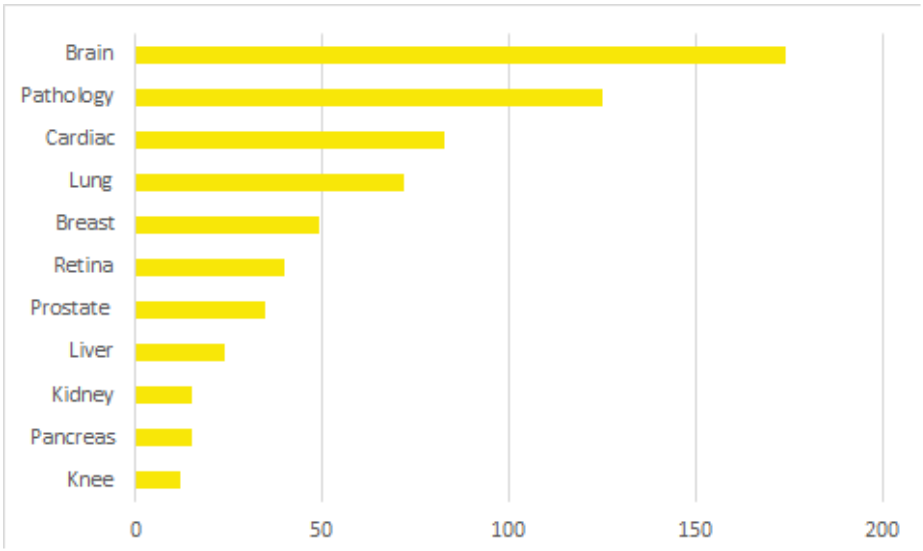


Figure 1.5: The distribution of research studies in the different body organ.

Computed Tomography (CT) or Computerized Axial Tomography (CAT) scan is a medical imaging method that combines multiple X-ray projections taken from different angles to produce detailed cross-sectional images of areas inside the body. CT scans provide 3-D views of certain parts of the body, such as soft tissues and bones.

CT is an accurate technique for the diagnosis of many abdominal diseases such as liver, lung, and pancreatic cancers. CT images allow doctors to measure and evaluate organs in the pelvis, chest, abdomen, cardiac tissue, cardiovascular

diseases, liver size and location of lesions, bone injuries, colon health, blood flow, brain infarction, and tumors, etc.

In this thesis, we applied brain CT scans for ischemic stroke lesion segmentation and abdominal CT images for liver lesion segmentation.

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging is a medical imaging technique that uses the strong magnetic field, magnetic field gradient, and radio waves to generate images of body organ and tissues. MRI has a wide range of applications in medical diagnoses, such as preoperative staging of rectal and prostate cancer.

MRI is an accurate technique for neurological cancers, as it has better resolution than CT and offers better visualization. The contrast provided between gray and white matter makes MRI the best choice for many conditions of the central nervous system, including dementia, cerebrovascular disease, Alzheimer's, and glioma disease.

MRI works by measuring the radio waves emitting by atoms subjected to a magnetic field. The appearance of tissue in an MRI depends on the tissues chemical composition and which particular MR sequence is employed. The most common of the sequence is T2-weighted MRI, in which tissues with more water or fat appear brighter due to their relatively high number of hydrogen atoms. In contrast, bone (as well as air) has low signal and appears dark on T2-weighted images. For brain MRIs, T1-weighted with gadolinium contrast enhancement (T1-Gd) and Fluid Attenuated Inversion Recovery (FLAIR) are commonly used sequences along with T2-weighted images. Determining which sequences to use for a given disorder or body part requires careful research or radiology expertise.

In the current thesis, we evaluated the proposed approaches using recent public MRI challenges such as BraTS 2013-2018, ACDC 2017, and HVSMR 2016.

Microscopic Light Imaging

Microscopy imaging or live cell imaging obtains a better understanding of biological function through the study of cellular dynamics. Generally, live cell microscopes keep cells alive during observation, and the lenses are commonly enclosed in a micro-cell incubator. Microscopic images that we study in this thesis consist

of 2D and 3D time-lapse video sequences of fluorescent counterstained nuclei or cells moving on top or immersed in a substrate, along with 2D Phase Contrast and Differential Interference Contrast (DIC) microscopy videos of cells moving on a flat substrate. The videos cover a wide range of cell types and quality (spatial and temporal resolution, noise levels, etc.).

1.2.2 Medical Image Visualization

Many applications for medical image analysis require visualization and analysis of two-dimensional (2D) or three-dimensional (3D) objects. Visualization is the process of exploring, transforming, and view data as images to gain understanding and insight into the data. Visualization techniques provide tools for medical reconstruction, diagnosis, and analysis of anatomic structure and function of the human body.

In this thesis, we developed tools for reading, writing, and transforming 3D medical images (i.e. .nhdr, .mha, .mhd, .nii, .nii.gz) into 2D images based on itk [13] libraries. Additionally, we release several useful large-scale C++ libraries, with a particular focus for working with image patches, image augmentation, and enhancement. Our code is publicly available [1].

1.2.3 Medical Image Preprocessing

Pre-processing is an important step to bring all subjects in similar distributions, to do this end we have different prepossessing policies depending image modality and dataset size. In MRI preprocessing step, we apply different normalization techniques, bias field correction, and image augmentation.

Normalization

Due to the nature of MRI, even images of the same patient on the same scanner at different can have different intensities. Normalization aims to remove some variation in the data (e.g., different subject pose or differences in image contrast, etc.) that is known and so simplify the detection of subtle differences we are

[1] https://github.com/HPI-DeepLearning/MedicalImagePreprocessing

interested in instead (e.g., the presence of a pathology). Here, we will go over the most common forms of normalization:

Normalization of voxel intensities This form of normalization is highly dependent on the imaging modality. Typical zero-mean, unit variance normalization is standard for qualitative images (e.g., weighted brain MR images, where the contrast is highly dependent on acquisition parameters, typically set by an expert). If we employ such statistical approaches, we use statistics from a single full volume, rather than an entire database.

Spatial normalization Normalizing for image orientation avoids that the model will have to learn all possible orientations, which largely reduces the amount of training images required (see the importance of header attributes to know what orientation an image is in). We additionally account for voxel spacing, which may vary between images, even when acquired from the same scanner. This can be done by resampling to an isotropic resolution: Further normalization includes medical image registration packages (e.g., MIRTK, etc.) and registers the images into the same space so that voxel locations between images correspond to each other. A typical step in analyzing structural brain MR images (e.g., T1-weighted MR images) is to register all images in the training database to a reference standard, such as a mean atlas (e.g., the MNI 305 atlas). Depending on the degrees of freedom of the registration method, this can also normalize for size (Affine registration) or shape (deformable registration). These techniques remove some information in the image such as shape and size that might be important for analysis (e.g., a large heart might be predictive of heart disease).

Data Augmentation

To accurately generalize deep learning model to unseen test cases, we augment training images by simulating a variation in the data aims to be robust and prevents over-fitting. Similarly to normalization methods, we distinguish between intensity and spatial augmentations.

Adding uniform/non-uniform noise to training images and random offset contrast to handle differences between images are examples of intensity augmentations which we applied in some experiments.

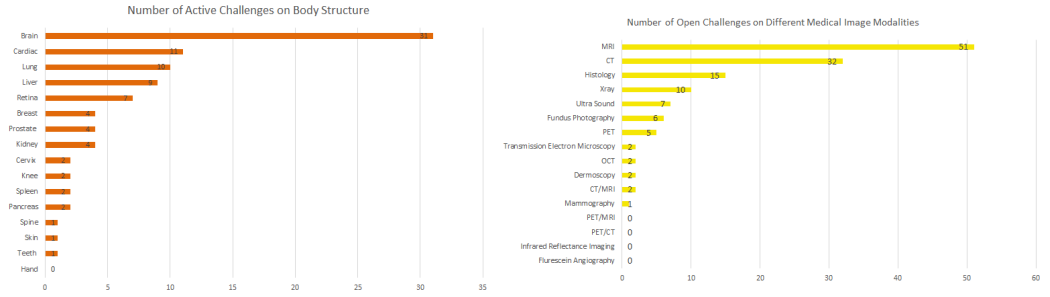


Figure 1.6: Number of open and active medical imaging challenges according to region of interest (application subjects) and imaging modality since 2007

Flipping the image tensor in directions on where to expect symmetry (e.g., a left/right flip on brain scans), random rotation along axes and random cropping scaling are examples of spatial augmentation.

Additionally, we generate diverse images based on training samples distribution using a generative adversarial network to oversample class minorities. We study the effect of oversampling minority classes by generating different image modalities using conditional GAN framework, implemented differently for 2D, 2D sequence, and 3D image generation.

1.2.4 Open Challenges

Applying deep learning algorithms for medical image analysis presents promising results in different application and open several unique challenges such as lack of large training data sets, having the acquisition of relevant these images, difficulties for annotating and labeling large dataset. Even when large datasets consist of different modalities and annotated by domain experts, label noise can be an important limiting factor in developing deep learning models. Commonly in computer vision, the noise in the labeling of images is typically relatively low. In medical image analysis, useful information does not depend on the images themselves. Physicians often leverage a wealth of data on patient history, age, demographics, and others to take a better decisions.

Figures 1.6 and 1.7 overview the most popular challenges which have been organized within the area of medical image analysis since 2007.

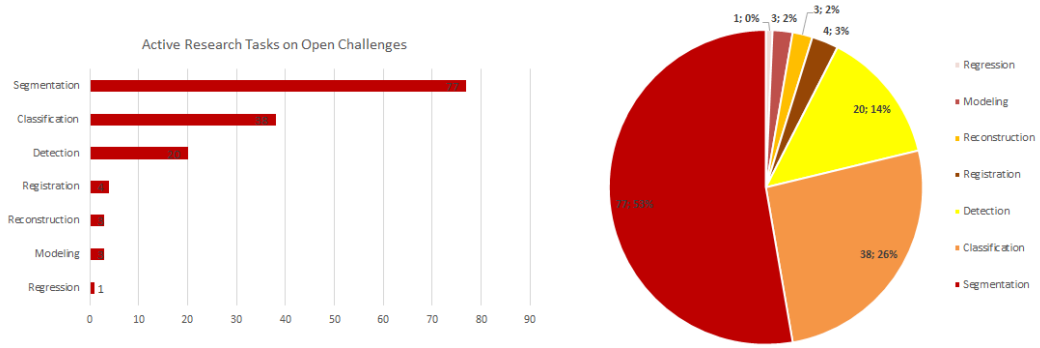


Figure 1.7: The distribution of open medical imaging challenges according to tasks

1.3 Nature of the Problem by Imbalanced Data

Traditional machine learning or deep learning models assume the number of samples in different classes are similar. However, in real life dataset, the class distribution is imbalanced, where some classes have a significantly higher number of examples in the training set than other classes. There are plenty of examples in domains such as medical diagnosis [14, 15], fraud detection [16], few-shot learning [17], autonomous driving [18], and others [19] where this issue is highly significant and the frequency of one class (e.g., cancer) can be 10000 times less than another class (e.g., healthy). It has been established that class imbalance can have a significant detrimental effect on training traditional classifiers [20] including multi-layer perceptrons [15] and convolutional neural networks [21].

The class imbalance problem affects convergence during the training phase and generalization of a model on the test set. The predicted results by these models have misclassification in minority classes where usually the minority classes are more important from the data mining perspective and they may carry important and useful knowledge.

The imbalanced class problem has been discussed in data mining [22] as well as machine learning literature [23]. Anand et al [24] analyses of learning from skewed data is related to convergence rates of back propagation-trained neural networks. Several surveys and papers captured recent advances in handling imbalanced class problem [25]. He and Ma [26] study important issues class imbalanced in sampling strategies, active learning, and streaming data. A book by Garca et al. [27]

discusses the effect of data preprocessing, among which a reasonable amount of space is dedicated to preparing, sampling, and cleaning imbalanced datasets. A more global review on learning from skewed data is proposed by Branco [28] and study on a more general issue of imbalanced predictive modeling. We summarize briefly some important parameters of class imbalanced dataset that influences machine learning algorithms such as data distribution, sample size, separability. We conclude this section with a brief overview of existing methods and evaluation measurement.

Imbalanced class distribution

The degree of imbalanced class distribution denotes by the ratio of the sample size of the small class to that of the majority class. In practical applications, the rate can be as drastic as 1:100, 1:1000, or even more significant. In this thesis, we study the effects of imbalanced class rate on segmentation and classification performances. Our finding indicates that a relatively balanced distribution usually attains a better result. However, at what imbalance degree the class distribution deteriorates the segmentation performance cannot be stated explicitly since other factors such as sample size and separability also affect performance.

Small sample size

Given a fixed imbalance rate, the sample size plays a crucial role in determining the *goodness* of a classification or segmentation model. In the case that the sample size is limited, uncovering regularities inherent in a small class is unreliable. In this thesis, we explore when the size of the training set increases, the large error rate caused by the imbalanced class distribution decreases in semantic segmentation. It expected when more data is used, relatively more information about the small class benefits the segmentation modeling, which becomes able to distinguish rare samples from the majority.

Seprateability

The difficulty in separating the small classes from the majority class is the critical issue of the small class problem. Assuming that there exist highly discriminative

patterns among each category, then not very sophisticated models are required to distinguish class objects. However, if patterns among each class are overlapping at different levels in some feature space, the discriminative model is hard to use.

1.3.1 Approaches for Handling Class Imbalanced

The significant difficulty of the class imbalance problem and its frequent occurrence in practical applications of machine learning have attracted research interests. Several research papers and holding workshops in ICML, AAAI conferences dedicated the importance of this problem. Research efforts study on two aspects of the class imbalance problem: (1) the nature of the class imbalance problem (e.g., in what domains do class imbalances most hinder the performance of a standard classifier? [29]); (2) the possible solutions in handling the class imbalance problem. Published solutions to the class imbalance problem divide into two groups as data-level and algorithm-level approaches.

Specifically, at the data-level objective is to re-balance the class distribution through resampling the data space. At the algorithm level, approaches try to adapt existing learning algorithms to strengthen learning with regards to the small class. In this thesis, we address the imbalanced class problem, both data-level and algorithm-level. Our approaches to skew data distribution and internal bias correction elaborate in Chapter 2. While, Chapters 3, 4, and 5 describe the solution in algorithm-level.

1.3.1.1 Methods on Data Level

Solutions at the data-level include many different forms of resampling techniques, such as randomly oversampling the small class, randomly undersampling the majority class, informatively oversampling the small class or undersampling of the majority class (e.g., using generative model), and combinations of the above techniques [30]. However, this oversampling and undersampling techniques often lead to remove some important samples or add redundant samples to the training set. Therefore, more advanced methods have proposed that try to maintain structures of groups and/or generate new data according to underlying distributions [31]. This family of algorithms also consists of solutions for cleaning overlapping objects and removing noisy examples that may negatively affect learners [32].

In this thesis, we synthesize minority class samples as well as missing image modalities according to underlying data distribution by adopting GAN based framework in two unsupervised and conditional supervised way. Moreover, at data-level, we introduce new patient-wise batch normalization motivated by the success of stratified sampling in precision medicine. We study the advantage of biased with true labels and complementary labels based on the transition probabilities for minority class in the application of classification and semantic segmentation. In the future, we plan to investigate the theoretical guarantee that the classifier learned with complementary labels converges to the optimal solution.

1.3.1.2 Methods on Algorithmic Level

Algorithm-level approaches focused on modifying existing learners to alleviate the bias towards majority class. These require good insight into the modified learning algorithm and precise identification of reasons for its failure in skewed mining distributions.

One-class learning is a straightforward solution which eliminates bias towards any group and concentrates on a single set of subjects. However, this technique needs some specific methods to use one-class learners for more complex problems. Another popular approach is cost-sensitive approaches which classifier is modified to incorporate varying penalty for each of the considered groups of examples. Here, the classifier assigns a higher cost to the less represented set of objects. The current thesis introduces two new costs, for instance, minority weighting.

Other methods include ensemble approaches combine same or different classifiers with improving generalization ability. The current thesis introduces three new ensemble architectures differs in type of ensemble, varied losses, and different image representation (2D, 2D sequential, and 3D).

1.3.2 Evaluation Measures

Evaluation metrics play a significant role in assessing the classification performance and guiding classifier modeling. Accuracy is a common evaluation metric which is not the best metric to use when evaluating imbalanced datasets as it can be very misleading. Metrics that can provide better insight include:

- Confusion Matrix shows correct predictions and types of incorrect predictions. Several measures can be derived using the confusion matrix:
  True Positive Rate: $TP_{rate} = TP/(TP + FN)$
  True Negative Rate: $TN_{rate} = TN/(TN + FP)$
  False Positive Rate: $FP_{rate} = FP/(TN + FP)$
  False Negative Rate: $FN_{rate} = FN/(TP + FN)$
  Positive Predictive Value: $PP_{value} = TP/(TP + FP)$
  Negative Predictive Value: $NP_{value} = TN/(TN + FN)$

- Precision presents the number of true positive divided by all positive predictions.Precision is also called Positive Predictive Value. It is a measure of a classifiers exactness. Low precision indicates a high number of false positives. $Precision = TP/(TP + FP)$

- Recall describes the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifiers completeness. Low recall indicates a high number of false negatives. $Recall = TP/(TP + FN)$

- F-Score represents a harmonic mean between recall and precision. The harmonic mean of two numbers tends to be closer to the smaller of the two. Hence, a high F-measure value ensures that both recall and precision are reasonably high. $F_1 = 2(precision \times recall)/(precision + recall)$

1.4 Contributions and Publications

This thesis introduces different novel generative adversarial frameworks that mitigate the imbalanced class problem for medical image segmentation and diseases classification. We evaluated our solution based on public benchmark an open challenge medical imaging in recent years. The results reported in this thesis mostly evaluated by challenge organizer using an online platform.

This thesis builds on several publications that earlier published in a peer-reviewed journal and presented international and scientific conferences. Besides, for the methods discussed in this thesis, we include open-sourced, accessible code on HPI-Deep Learning Github [1]. We release several useful C++ libraries, with particular focus to pre-process image patches.

The contributions of this thesis are as follows:

Internal Bias Correction by Synthesis Minority Sampling
We demonstrate that synthesize certain types of medical images either from noise alone as well as from prior knowledge by setting condition image data are able to mitigate imbalanced class problem and missing medical modalities. We adopt GAN framework as unsupervised setting for 2D, or 3D synthetic diverse image generation on minority classes. Then, we study the advantage of generated samples together with training data in balanced dataset to deep learning model. In Chapter 2, we study the effects of synthesized data rather than classical data augmentation for handling imbalanced data using the same deep learning framework. Hence, the diversity gained from synthesized images can introduce more variability into the training set and considerably improve classification results. The results are discussed according to quantitative, qualitative evaluation, and also, we compute the gradient of output category regarding input images to interpret the decision.

Bias Correction by Stratified Batch-Normalization
We introduce patient-wise batch normalization motivated by the succeed of stratified sampling in personalized medicine. The patient-wise batch normalization is able to skew data distribution by normalizing batches within homogeneous and among heterogeneous of samples. The proposed method enforces a deep network to maintain on a similar distribution where the patient-wise batch normalization is able to scale the whole underlying probability density function described in Chapter 2.

Handling Imbalanced by Cost Weighted Instance Learning

[1] https://github.com/HPI-DeepLearning/MedicalImagePreprocessing

In Chapter 3, we present two new cost-sensitive learning losses that modifying existing learners to alleviate their bias towards majority groups. As a first approach, we assign a higher cost to the less represented set of objects and boost its importance during the learning process. In the second approach, we weighted loss function to attenuate the effect of majority class frequency. Then, we show the application of weighting losses on handling the imbalanced class problem for semantic segmentation and diseases classification.

Learning Imbalanced with Class Expert Ensemble

In Chapter 4, we demonstrate that an ensemble of multiple models as a single consensus model can efficiently handle the imbalanced class problem. We introduce three different ensemble models differs from architecture, loss, and type of ensemble based on the generative adversarial network. We explain multi-objective optimization guarantee converges to the optimal solution. Then, we show the application for highly imbalanced semantic segmentation task.

Learning Imbalanced with Mutual Information

We explore the mutual information shared between independent generators helpful to mitigate the imbalanced class problem. We implemented different GAN based architecture and examined the impact of mutual learning for semantic segmentation of imbalanced data. The details of architecture, training procedure, and experiments are demonstrated in Chapter 5.

The list of selected publications to this thesis includes the following:

- Mina Rezaei, Haojin Yang, Christoph Meinel: Recurrent Generative Adversarial Network for Learning Imbalanced Medical Image Semantic Segmentation. Proceeding by Journal of Multimedia Tools and Application, special issues on Computer-Aided Radiology and Diagnosis.

- Mina Rezaei, Haojin Yang, Konstantine Harmuth, Christoph Meinel: Conditional Generative Refinement Adversarial Networks for Unbalanced Medical Image Semantic Segmentation. Proceeding by IEEE Winter Conference on Application Computer Vision WACV 2019.

1. LEARNING REPRESENTATION FROM IMBALANCED MEDICAL IMAGING

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Learning Imbalanced Semantic Segmentation through Cross-Domain Relations of Multi-Agent Generative Adversarial Networks. Proceeding of SPIE Medical Imaging - Computer Aided Diagnosis (SPIE Medical Imaging 2019 - **Best Student Paper Award**)

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Multi-Task Generative Adversarial Network for Handling Imbalanced Clinical Data. Accepted and presented in Machine Learning for Health Workshop at Advances in Neural Information Processing Systems 2018 (ML4H)

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Generative Adversarial Framework for Learning Multiple Clinical Tasks. Proceeding by Digital Image Computing: Techniques and Applications (DICTA 2018)

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: voxel-GAN: Adversarial Framework for Learning Imbalanced Brain Tumor Segmentation. Proceeding by BrainLes@MICCAI 2018.

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Instance Tumor Segmentation using Multitask Convolution Neural Network. Proceeding by IEEE Joint Conference on Neural Networks (IJCNN 2018)

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Automatic Cardiac MRI Segmentation via Context-aware Recurrent Generative Adversarial Neural Network. Accepted and presented in Computer Assisted Radiology and Surgery (CARS 2018)

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Whole Heart and Great Vessel Segmentation with Context-aware of Generative Adversarial Networks.Proceeding by Bildverarbeitung fr die Medizin 2018.

- ☐ Mina Rezaei, Haojin Yang, Christoph Meinel: Deep Neural Network with l2-norm Unit for Brain Lesions Detection. Proceeding by 24th International Conference Of Neural information Processing (ICONIP 2017)

- ☐ Mina Rezaei, Konstantin Harmuth, Willi Gierke, Thomas Kellermeier, Martin Fischer, Haojin Yang, Christoph Meinel: A Conditional Adversarial Network for Semantic Segmentation of Brain Tumor. Proceeding by BrainLes@MICCAI 2017.

- ☐ Mina Rezaei, Hiroyuki Yoshida, Haojin Yang, Christoph Meinel: Generative Adversarial Ensemble Discriminators for Handling Class Imbalanced Problem in Medical Image Segmentation, Under Review 2019.

- ☐ Mina Rezaei, Hiroyuki Yoshida, Haojin Yang, Christoph Lippert, Christoph Meinel: Deep Mutual GAN: An Adversarial Frameworks for Handling Class Imbalanced Problem in Medical Image Segmentation, ready to submit.

1.5 Thesis Structure

The outline of the thesis is as follows.

After the introduction presented in this chapter, Chapter 2 presents the data-level contributions (Section 2.1) to mitigate class imbalanced problem in task of medical image classification and semantic segmentation. The detail architecture of GAN-based framework, the dataset, evaluation, and comparison are described in Section 2.2. Similar GAN-based architecture and recent data-level approaches for internal bias correction are described in Section 2.3. In Section 2.4, summary and future direction for data-level approaches are discussed.

Chapter 3, introduces two novel cost proportional weighting to mitigate class imbalanced problem for semantic segmentation as well as multi-task GAN architecture. Section 3.1 reviews of the related GAN-based architecture for medical image segmentation and multi-task learning as well as cost sensitive losses for mitigating imbalanced data. The detail of network architecture, implementation, evaluation, and comparison with other participant in the challenges are described in Section 3.3.

Chapter 4 presents the three different proposed ensemble generative adversarial networks for handling class imbalanced problem (Section 4.2). The detail of experimental setting, implementation, evaluation, and comparison based on open benchmark dataset are discussed in Section 4.3 and Section 4.3.4.

1. LEARNING REPRESENTATION FROM IMBALANCED MEDICAL IMAGING

Last chapter, address learning representation of imbalanced data using mutual information shared between independent models. Section 5.2 introduces the technical background, Section 5.2 formulates the proposed deep mutual GANs including couple of generators and couple of discriminators, and Section 5.2.2 explains two network architectures. The extension of deep mutual GAN for more than four agents are discussed in Section 5.2.3. The detail of application on semantic segmentation, implementation of end-to-end deep learning architectures, evaluation, and comparison are described in Section 5.3. In Section 5.2.3, we make overall conclusions, discuss the application of the algorithm-level solution and suggest directions for future research.

Chapter 2

Novel Approaches for Internal Bias Correction on Imbalanced Data

Many researchers address learning from the imbalanced sample on the data-level that concentrate on modifying the training set to make it suitable for standard learning algorithms. In order to balance data distribution, we may notice approaches that generate new objects for minority class (oversampling) [33, 34] and that remove examples from the majority class (undersampling) [35]. However, these approaches often lead to remove some important samples or add redundant samples to the training set. Therefore, more advanced methods are proposed, including variational autoencoder (VAE) and generative adversarial networks (GANs) that try to maintain structures of groups and generate new data according to underlying distributions. This family of algorithms also consists of solutions for cleaning overlapping objects and removing noisy examples that may negatively affect learners [36]. In this regard, we synthesize minority class samples using deep convolutional GAN and generate new samples 2D or 3D and make training sample balance and suitable for a standard deep learning model.

Other techniques include incremental rectification of mini-batches for training deep neural network [37, 38]. Recently Dong et al. [37] introduce batch-wise

mining to tackle with imbalanced of minority class incremental rectification using a deep convolutional neural network. Plotz et al. [38] approach new batch sampling based on K nearest neighbors. Here, we study internal bias correction on the batch level motivated by the success of stratified sampling in personalized medicine. Therefore, we propose a patient-wise batch normalization technique which enforces a deep network to maintain on a similar distribution. As shown by Ioffe et al. [39], batch normalization is proposed to alleviate the internal covariate shift by normalizing layer outputs for each training mini-batch with respect to its very own statistics, specifically mean and variance. We observed better trade-off between precision and recall compared to when conventional batch sampling. This is due to the fact that the variability within the subgroups is lower compared to the variations when dealing with the entire population at large.

The rest of the chapter is organized as follows: the next Section 2.1 explains the proposed synthesis minority class using unsupervised and conditional GAN (see Section 2.1.1) and the patient-wise mini-batch sampling (see Section 2.1.2) for balancing data distribution. The detailed experimental results are presented in Section 2.2. The overview of recent relevant methods for handling imbalanced training data related to our contributions is described in Section 2.3. We make an overall conclusion on data-level approaches and explain the complementary label for future research in Section 2.4.

2.1 Internal Bias Correction

Here, we explain our contributions on internal bias correction and balancing data distribution by approaching synthesis class minority samples and patient-wise batch normalization techniques. Then, we applied our data-level approaches for medical image classification to diagnosis diseases, or semantic segmentation using recent state-of-the-art deep learning model and generative adversarial networks.

2.1.1 Biased with Synthesis Minority Oversampling

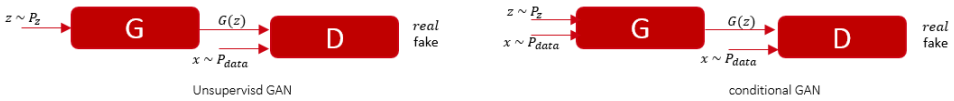


Figure 2.1: Comparison of unsupervised GAN and conditional used for synthetic minority oversampling.

Generative adversarial networks (GANs) have been proposed widely for image generation. We can group the GAN-based approaches into two categories: unsupervised generative framework and conditional GANs. Unsupervised GAN maps random noise to synthetic, realistically looking images following the training data distribution while the conditional GAN framework has a condition on prior knowledge both the generator and the discriminator rather than noise alone. Figure 2.1 shows differ between unsupervised GAN and conditional GAN framework.

In this thesis, we adopt GAN framework to synthesize certain types of medical images either from the noise alone as well as from prior knowledge by setting condition image data for handling the imbalanced class problem and missing medical image modalities.

As an unsupervised setting, we implemented different GAN framework for 2D or 3D synthetic diverse image generation on minority class examples. Then, the generated images, together with training samples as a balanced dataset, are fed to deep learning models. We study the effects of balancing dataset by synthesized data rather than classical data augmentation using the same deep learning framework. Hence, the diversity gained from synthesized images can introduce more variability into the training set and considerably improve classification results. In continue, we describe our GAN framework for generating minority class example.

Generating Synthetic Brain Images using Unsupervised GAN

Generative adversarial network categories as a class of a generative model which aims to implicitly learn the data distribution p_{data} from a set of samples (e.g., images) to further generate new examples from random noise p_z drawn from the learned distribution. Deep convolutional GAN (DCGAN) consists of two deep CNNs models that are trained separately and simultaneously, as depicted in the left side of Figure 2.1.

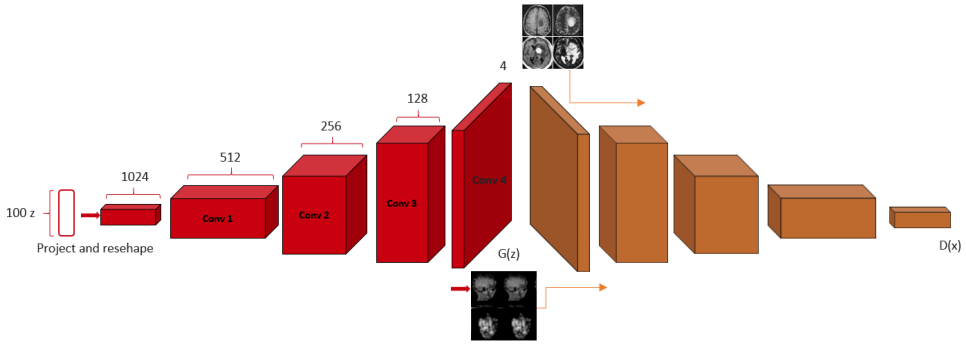


Figure 2.2: Unsupervised deep convolutional GAN architecture for synthetic minority oversampling

The generator takes random input vector z from a known simple distribution p_z and maps $G(z)$ to the image space of distribution p_g. A sample x is input to the discriminator which outputs $D(x)$, its probability of being a real sample. During training, the generator improves in its ability to synthesize more realistic images while the discriminator improves in its ability to distinguish the real from the synthesized images. The GAN objective function is a two-player mini-max game like Eq.(2.1).

$$\min_G \max_D V(D,G) = E_y[logD(y)] + E_z[log(1 - D(G(z)))] \tag{2.1}$$

Figure 2.2 shows the DCGAN architecture which we used for minority class image generation.

Balancing Data Distribution with conditional GAN for Synthesis Missing MRI Modalities

In different architecture, we develop a condition GAN framework to synthesize 2D sequence medical images and balance the data distribution regarding missing image modalities. The generator and the discriminator condition by prior knowledge to generate more realistic and high-resolution images from missing image modalities from imbalanced class categories. Here, we use least absolute deviations (see Eq.2.3) as reconstruction loss in additionally with an adversarial loss to improve the realism of the synthetic images.

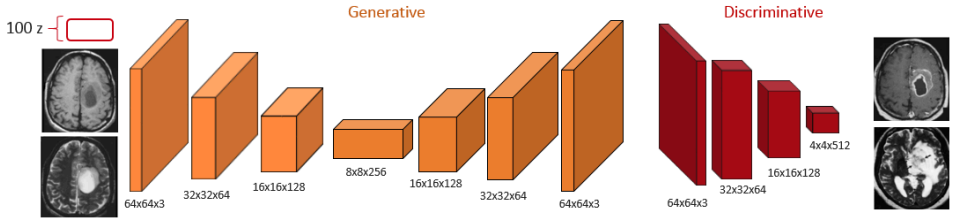


Figure 2.3: Deep conditional GAN architecture for synthetic minority oversampling regarding missing modality

Unlike previous conditional GANs [11, 40, 41, 42, 43]; in our proposed method, a generative model learns mapping from a given sequence of 2D multimodal MR images x_i to a sequence missing modalities o_i; $G : \{x_i, z\} \rightarrow \{o_i\}$ (where i refers to 2D slice index between 1 and 155 from a total 155 slices acquired from each patient). We utilize bidirectional LSTM to pass the temporal consistency between 2D slices. Our network can learn representations from previous and future slices, which results in context-aware and eliminate ambiguity. The training procedure for generating missing image modalities is similar to two-player mini-max game, as shown in Eq.(2.2). Similar to the image-to-image translation [40], we translate or synthesize multi-modal images (e.g., MRI T1, T2, T1c) to one missing-modality (e.g., MRI Flair) or map one image (e.g., MRI T1) to different image-modalities (e.g., MRI Flair, T2, T1c).

$$\mathcal{L}_{adv} \leftarrow \min_{G} \max_{D} V(D,G) = E_{x,o_{modality}}[logD(x, o_{modality})] + E_{x,z}[log(1 - D(x, G(x,z)))] \tag{2.2}$$

$$\mathcal{L}_{lad} = \sum_{i=1}^{n} |D(x_i, o_{modality}) - G(x_i, z)| \tag{2.3}$$

where $G(x_i, z)$ is synthesized images by generator while $D(x_i, o_{modality})$ is actual missing image modality. The final objective function then calculated by adding ℓ_{lad} to ℓ_{adv}.

As shown by Figure 2.3, the generator is autoencoder that takes one or three available MRI modalities and able to generate three or one missing modalities by getting feedback from the fully convolutional discriminator networks.

In following, we describe the deep convolutional network that we used for classification on an unbalanced dataset and balanced dataset by unsupervised GAN, conditional-GAN, and classical data augmentation.

Deep Convolutions Network for Brain Abnormality Classification
We apply recent successful deep residual network, winner of the classification task of ILSVRC-2016, for brain abnormality classification. The ResNet [44] is well-known due to its depth (152 layers) and the introduction of residual blocks. The residual blocks address the problem of training a really deep architecture by introducing identity skip connections so that layers can copy their inputs to the next layer. The main idea is to ensure the next layer learns something new and different from what the input has already encoded (since it is provided with both the output of the previous layer and its unchanged input). Moreover, this type of connections helps to overcome the vanishing gradients problem.

We train small ResNet with 18 layers, comprises by five convolutional layers, max pooling, fully connected, and softmax as the loss function. Our classification network takes 2D images with three channels, while each channel contains a grayscale copy with the same size and same plane from various MRI modalities with respective class label l=0,1,..,4. Each grayscale copy extracted from T1, T1c, and FLAIR of the same MRI categories has been mapped to the Red, Green and Blue channels of a standard image container, respectively. In the experiments, we observed the effect of oversampling minority class generated by different GAN architecture compared to classical data augmentation as well as imbalanced training data.

2.1.2 Biased with Patient-wise mini-batch Sampling

Several popular techniques are developed for normalization, such as batch normalization [45], and max norm constraints [46], with the core idea of shifting the inputs to a zero mean and unit variance. The input data is normalized before applying non-linearity to prevent the input from saturating extreme non-linearity. As described by Ioff et al. [45], batch normalization improve the overall

optimization and gradient issues. In many cases, initial weights have large deviance from true weights, delaying the convergence during training. Batch norm reduces the influence of weight deviance by normalizing the gradients this speed up the training.

Later, Ioffe [47] proposed a new Batch Renormalization since the conventional batch normalization is not well suited to small training set consists of different samples. Batch Renormalization [47] replaces batchnorm to ensure that the outputs computed by the model are dependent only on the individual examples and not the entire dataset, during both training and testing. In medical imaging often we have small dataset includes different samples, motivated by this we reformulated batch normalization by stratified sampling.

Stratified batch sampling shows successful results in personalized medicine [48] and statistic [49] when sub-populations within an overall population vary. Stratified sampling can reduce variance [50] through sampling each sub-population (stratum) independently where the strata are constructed within homogeneous among heterogeneous. Nguyan et al. [51] provide theoretical proof of stratified sampling on different deep models to reach the variance-optimal.

Similar to the concept of stratified sampling, we initially normalized the inputs where the mean and variance are computed on a specific patient from the same acquisition plane (Sagittal, Coronal, and Axial) and from all available image modalities (e.g., T1, T1-contrast, T2, Flair in the BraTS benchmark). In this regard, the deviance get increasingly large, and the back-propagation step needs to account for these large deviance which this restrict us from using a small learning rate to prevent gradient explosion. Each stratum enforces to mini-batch consists of different patches among similar samples (patients). For example, the mini-batch with 128 images includes the same patient images and four available modalities from the same acquisition plane. Algorithm 1 shows how to compute normalization at each mini-batch by proposed patient-wise batch-norm technique.

2.2 Experiments

Algorithm 1: Patient-wise mini-batch normalization. (Here we use three site MR images S, C, A Sagittal, Coronal, and Axial respectively. i and n respectively refer to a number of 2D slices and number of patient e.g. $0 < i \leq 155$, n=230 in BraTS). L, N_h, and n_h refer to the number of strata, number of units in each stratum h, and the number of samples taken from stratum h respectively.

Input : Values of x over a mini-batch: $\beta = x_1, x_2, ..., x_{155}$
Parameters to be learned:γ, β

Output: $y_i = BN_{\gamma,\beta}(x_i)$

for $Patient : P_1, P_2, ..., P_n$ **do**
  for $AcquisitionPlane : S_i, C_i, A_i$ **do**
    for $Image \quad Modalities : T1, T2, T1c, Flair$ **do**
      $\tau_h \leftarrow N_h x_i$
      $\tau_{st} \leftarrow \sum_{h=1}^{L} \tau_h$
      $\mu_{st} \leftarrow \frac{1}{m}\tau_{st}$
      $\sigma^2_{st} \leftarrow \frac{1}{m}\sum_{i=1}^{n}(x_i - \mu_{st})^2$
      $\hat{x_i} \leftarrow \frac{x_i - \mu_x}{\sqrt[2]{\sigma^2_x + \varepsilon}}$
      $y_i \leftarrow \gamma\hat{x_i} + \beta = BN_{\gamma,st}(x_i)$
    end
  end
end

To evaluate the performance of our approaches on imbalanced data and compared them with state-of-the-art methods, we trained recent popular annotated medical imaging benchmarks on the task of semantic segmentation and classification.

2.2.1 Experiments on Balanced with Synthesized GAN Samples

We tested generated minority samples using unsupervised and conditional GANs for brain disease diagnosis with different imbalanced ratio. In this experiment, we applied real patient data from five popular benchmarks to evaluate the effect of synthetic images as balancing solution on the training set. For the classification task, we compiled 1500 MRI images with the label of healthy, tumor-HGG, tumor-LGG, Alzheimer, and multiple sclerosis. We consider 20% of the data for testing and 80% for training

IXI dataset[1] contains 600 MRI images from normal, healthy subjects. The MRI image acquisition protocol for each subject includes six modalities, from which we have used T1, T2, PD, MRA images. The first column of Figure 2.4 shows the healthy brain images from IXI dataset in the sagittal, coronal, and axial sections. The BraTS2016 benchmark [2] prepared the data in two part of high and low grade glioma (HGG/LGG). All images have been aligned to the same anatomical template and interpolated to 1 mm voxel resolution. The training dataset consists of 220 HGG and 108 LGG MRI images, which for each patient T1, T1contrast, T2, FLAIR, and ground truth labeled by medical experts have been provided. Alzheimer disease dataset[2] comes from the Open Access Series of Imaging Studies (OASIS). The dataset consists of a cross-sectional collection of 416 subjects aged from 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans were obtained in single scan sessions. 18 MRI images with multiple sclerosis from ISBI challenges 2008 [52] have also been applied in the classification task. ISLES benchmark 2016 [53] (Ischemic Stroke Lesion Segmentation) comes from MICCAI challenge in two-part, by which we used only SPES dataset with 30 brain images provided in seven modalities in our task.

[1] http://brain-development.org/ixi-dataset/
[2] http://www.oasis-brains.org/

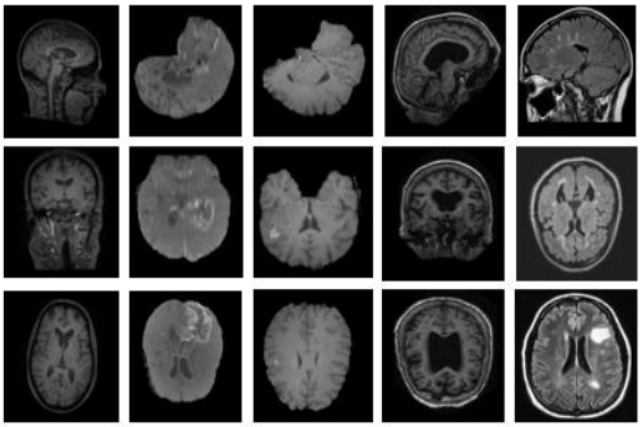


Figure 2.4: The five different categories of brain MR images. The first column shows healthy brain in sagittal, coronal, and axial plane. The 2nd and 3rd columns show high and low grad glioma, while 4th and 5th columns present some brain images on Alzheimer and multiple sclerosis.

Because the MRI volumes in the BraTS 2016 and ISLES 2015 datasets do not possess an isotropic resolution, we prepared 2D slices in sagittal, axial, and coronal view. As mentioned by Havaei et al. [54], brain imaging data are rarely balanced even in one category due to the different size of the lesion in the brain. For example, the volume of a stroke is rarely more than 1% of the entire brain, and a tumor (even large glioblastomas) never occupies more than 4% of the brain. Training a deep network with imbalanced data often leads to very low true positive rate since the system gets to be biased towards the one class that is over-represented. To overcome this problem, we have chosen 2D slice of MRI with lesions. In our unsupervised GAN setting, the deep generative model takes only random noise and tries to capture the distribution of 2D slices with lesion while discriminative network classifies whether generated images look realistic or fake. To evaluate the synthesized images, we prepared the evaluation based on FID, a distance metric for GANs framework.

Fréchet Inception Distance

We used the FID [55] to calculate the distance between generated sample distribution p_g and true sample distribution p_d based on trained generative model G. We can sample from p_g many times, but directly, we are not able to evaluate p_g.

Table 2.1: The minimum FID obtained across 100 epochs for generating five categories of brain images by conditional GAN and unsupervised GAN framework.

Methods	Healthy	Alzheimer	HGG	LGG	MS
Unsupervised-GAN	4.816491	8.657932	3.980030	2.04520463	5.10204146
cGAN	0.888543	2.697015	1.107877	0.96701305	1.12549001

Table 2.2: The achieved average FID by unsupervised and conditional GAN which obtained across 100 epoch.

Methods	Healthy	Alzheimer	HGG	LGG	MS
Unsupervised-GAN	32.657932	49.9587135	51.0458763	47.5124854	63.7718045
cGAN	26.888543	30.1078707	28.5721066	35.5149001	48.6875241

In this regard, Fréchet Inception Distance [55] is suggested a way for evaluation of such a model. The distance is calculated as follows Eq. (2.4):

$$FID(t,p) = \parallel \mu_t - \mu_p \parallel^2 + Tr(cov(t) + cov(p) - 2(cov(t)cov(p))^{\frac{1}{2}}) \tag{2.4}$$

Where t and p respectively refer to the ground truth (or real images) and predicted mask (or generator output), μ and cov indicate mean and covariance of a multi-variate Gaussian produced from the embedding of the last pooling layer of the Inception-v3 model [56].

Table 2.1 shows the minimum, FID obtained by G at the end epochs 100 for generated brain four categories of brain images. Lower values indicate more similarity between the training data and synthesized sample by our unsupervised GAN. Based on Table 2.2, we observe the average values of FID after each 10 epochs across 100 epochs in total.

A subset of the generated samples produced by the unsupervised-GAN reported in Table 2.2 are shown in Figure 2.5, where we observe high variety and realism across all generated sets.

Here, we used ResNet-18 consists of five convolutional layers and two residual blocks in each, two dense layers, and softmax as the loss function. As we

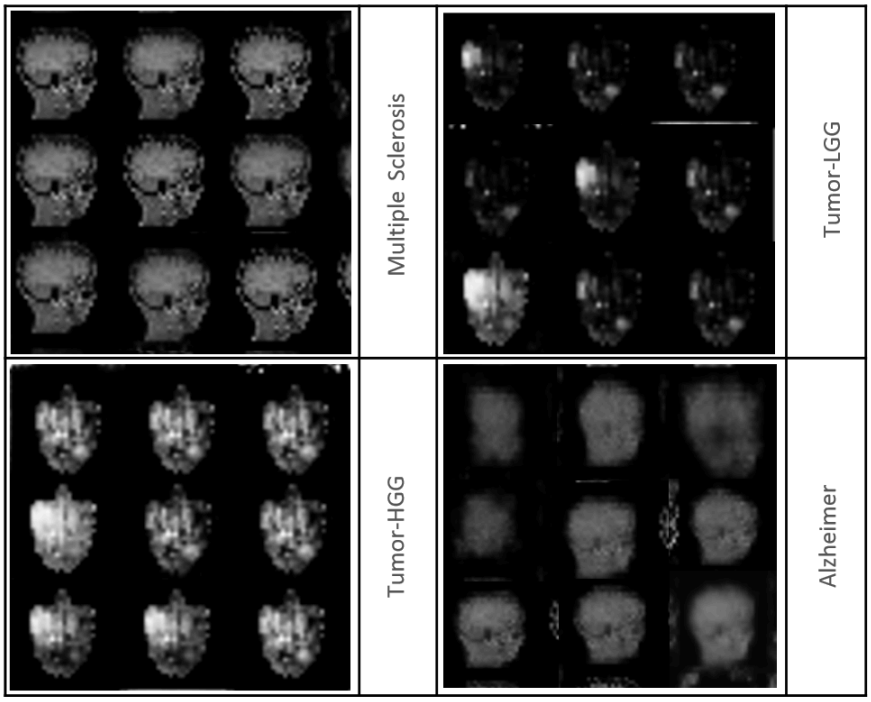


Figure 2.5: The generated samples from four different categories of brain MRI.

expected the synthetic augmentation compared to the classic augmentation provides significant improvement. Figures 2.7,2.8, and 2.9 demonstrate and compare the confusion matrix of the classification result on imbalanced data and balanced training set with classical data augmentation and the generative unsupervised synthesized framework respectively.

In different experiments, we developed conditional recurrent-GAN to generate missing modalities. To this end, we trained the recurrent-GAN separately with some available MRI or CT modalities from BraTS 2016 and ISLES 2015 dataset respectively. Here, we aim to observe whether the synthesized samples are able to improve classification results as a real example. To this end, we trained conditional recurrent GAN to generate T1c image modality where the condition is on T1c, T2, and Flair image modalities. Figure 2.6, fourth column shows the generated images by recurrent conditional GAN and Table 2.3 first column provides a quantitative evaluation based on mean and minimum FID. As second experiment, the recurrent GAN is trained with T2 and T1c and the generator learns to synthesize Flair and T2. The results are reported in third and fourth

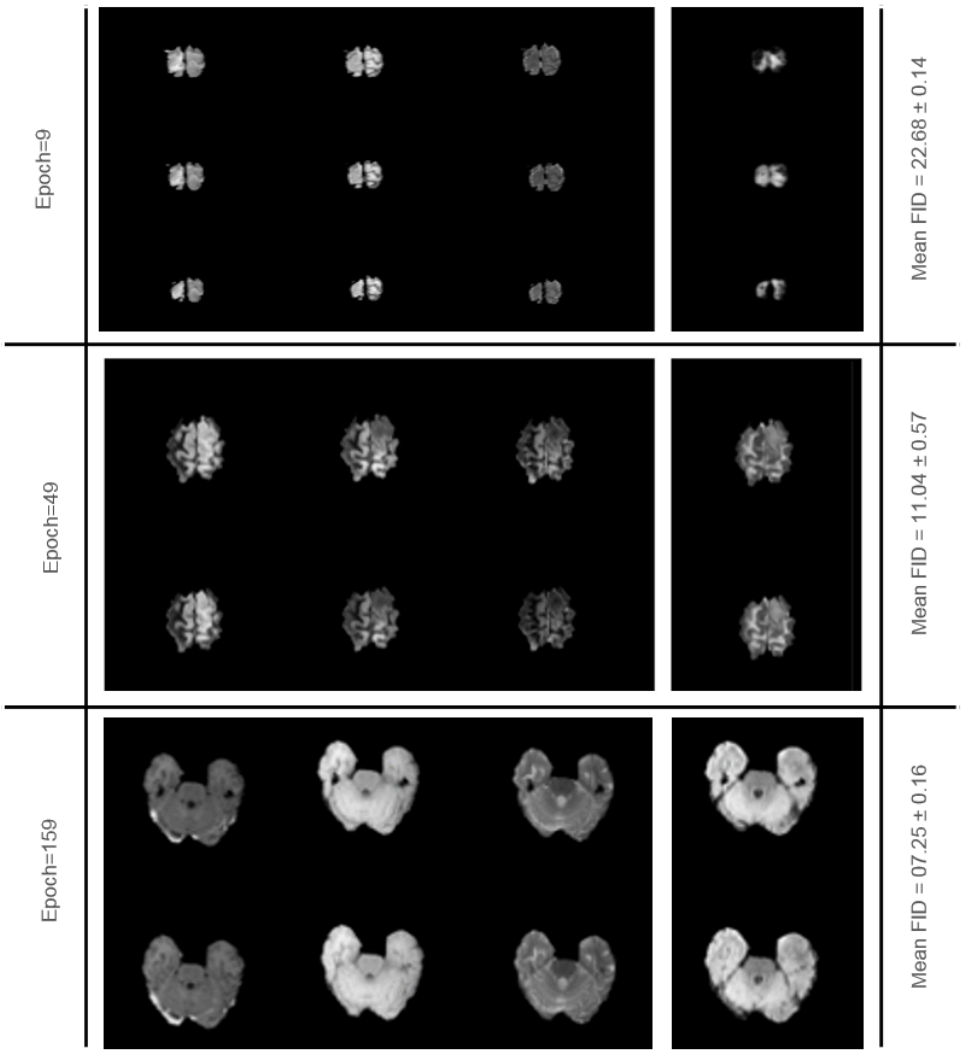


Figure 2.6: The synthesized sample by conditional GAN on BraTS dataset regarding missing modality. Column fifth generated by G in test time with condition input of columns 2-4.

column of Table 2.3. The last column of Table 2.3 shows our results based on ISLES 2015 dataset for generating DWI modality which T2, T1c, and Flair are the conditions.

It is expected that more generalized features could be able to learned from multiple modalities, and the testing accuracy based on more generalized features should be gained. The classification results from Table 2.4 proved our assumption, where better detection results were achieved by increasing the data modalities in

Table 2.3: The first row represents the minimum FID obtained across 100 epochs. The second row shows the mean calculated on each 10 achieved by recurrent-GAN for generating 2D sequence of missing image modality.

FID FID	T1c BraTS-1	Flair BraTS-2	T1c BraTS-2	DWI ISLES
min FID	0.96491	0.657932	0.980030	1.0458763
mean FID	30.888543	21.697015	19.107877	41.9587135

Table 2.4: Dice Similarity Coefficient (DSC) results (for brain lesions diagnosis) on the BraTS2016 and ISLES2016 dataset by generated samples from recurrent-GAN. F/D column means the FLAIR modality in BraTS dataset and DWI modality in ISLES dataset.

T1	T1c	T2	F/D	Dice-BraTS16	Dice-ISLES
x	-	-	-	61.8 %	42%
-	x	-	-	39.76 %	27%
-	-	x	-	36.7 %	39.98%
-	-	-	x	73.38 %	50.71%
-	x	x	x	81.53 %	54.23%
x	x	-	x	82.6 %	54.67%
x	-	x	x	85.19 %	53.09%
x	x	x	-	84.73 %	54.7%
x	x	x	x	89.53 %	78.17%

the model training.

Table 2.5 and Table 2.6 provide classification comparison based on imbalanced dataset and balanced with generated samples or classical data augmentation. Here we applied horizontal and ventricle flipping, multiple scaling, and [-10, +10] image rotation only in abnormal classes.

Saliency Map Visualization

Visualization of weights can present useful information about what the trained network is learning. Specially, in medical diagnosis it would be irresponsible to trust prediction of black-box system. Therefore, we used class saliency [58] to visualize the trained ResNet18 weights in prediction time. The goal of saliency is to find the pixels of an image which contribute most towards a particular

Table 2.5: Evaluation result of the classification of imbalanced and balanced dataset with generating data augmentation and SMOTE-GAN.

	Dice(imbalanced)	Dice(with data augmentation)	Dice(with SMOTE-GAN)
BraTS16	72%	83.53%	89.53%
ISLES16	59.65%	71.87%	78.17%

Table 2.6: Brain lesions classification performance of the ResNet architecture. The involved classes include healthy, tumor-HGG, tumor-LGG, Alzheimer, and multiple sclerosis.

	Total MRI	Accuracy	Sensitivity	Specificity	Recall	Kappa
Our method	1500	96.308%	0.91	0.87	87.65	0.92
Justin S et al. [57]	191	91.43%	-	-	-	-

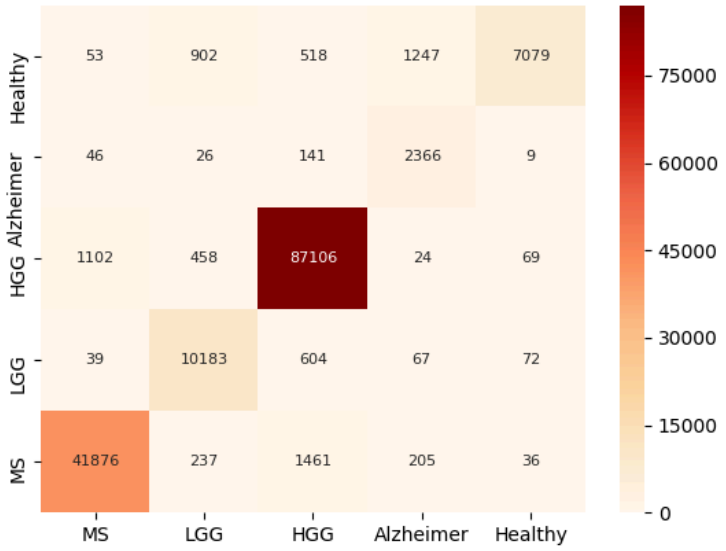


Figure 2.7: Confusion matrix on classification of five categories brain diseases predicted by ResNet18 which the network trained on highly imbalanced dataset.

classification. Figures 2.10,2.11 show which pixels in the images that are most important in the trained ResNet to classify it as an image of high grade glioma of Tumor (HGG). Here, we take derivative of the class score S_c with respect to the input image I, and evaluate at test time I_0; mathematically: $\frac{\partial S_c}{\partial I}|I$ where

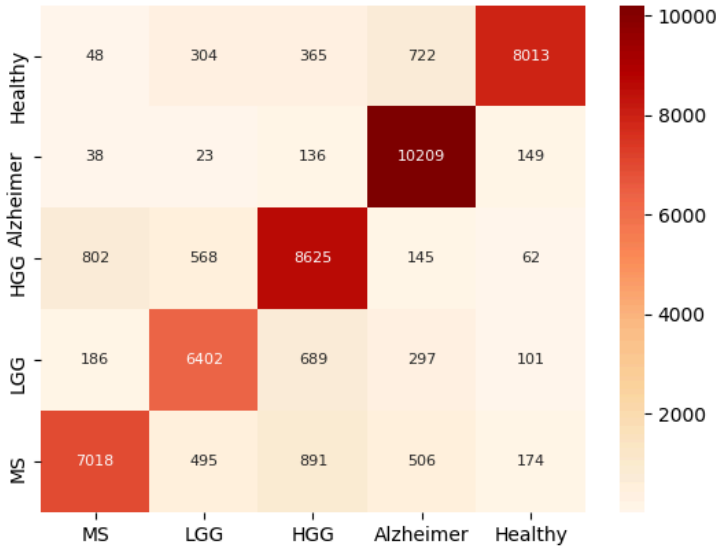


Figure 2.8: Classification results obtained by ResNet18 trained on imbalanced dataset by classical data augmentation

this derivative gives us a scalar quantity for each of the pixels in the image. Let w_{ij}^c, be the quantity at location (i, j) when weve used the score for class$c$. We can take the magnitude of these values and then normalize them to get a class saliency map M_c over the image.

$$M_c(i, j) = \frac{w_{ij}}{\sum_{ij} w_{ij}} \tag{2.5}$$

We compute the gradient of output category with respect to input image. This should tell us how output category value changes with respect to a small change in input image pixels. All the positive values in the gradients tell us that a small change to that pixel will increase the output value. Hence, visualizing these gradients, which are the same shape as the image should provide some intuition of attention.

The idea behind saliency is pretty simple in hindsight. We compute the gradient of output category with respect to input image. This should tell us how the output value changes with respect to a small change in inputs. We can use

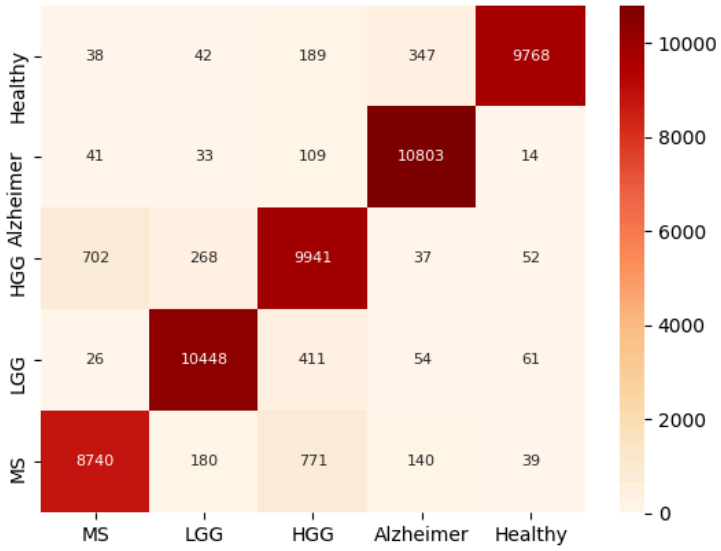


Figure 2.9: Brain disease classification results achieved by ResNet18. The network trained on balanced dataset with synthesized samples generated by unsupervised GAN.

these gradients to highlight input regions that cause the most change in the output. Intuitively this should highlight salient image regions that most contribute towards the output.

To visualize what contributed to the predicted output, we want to consider gradients that have very low positive or negative values.

Moreover, the image saliency map (the top, left side image in Figure 2.10) can be used for localizing an object of interest (in the above example, we can see where the dog in the image is), and segment it out with the help of a segmentation algorithm. Note that the classification model is not trained with object locations; its only given (image, category) pairs, but learns to localize: this is called weakly-supervised object localization.

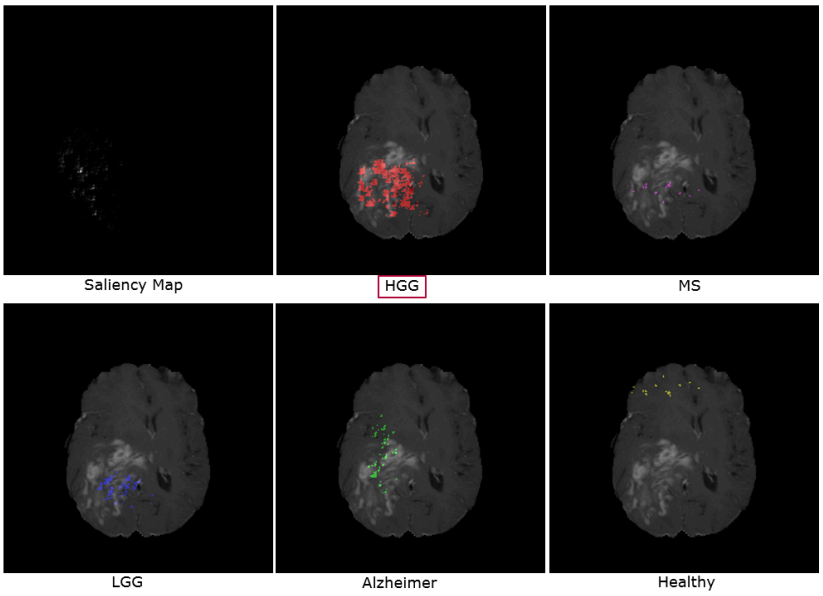


Figure 2.10: Saliency map visualization of ResNet18 which shows pixels that are most important for the image being classified as a HGG glioma.

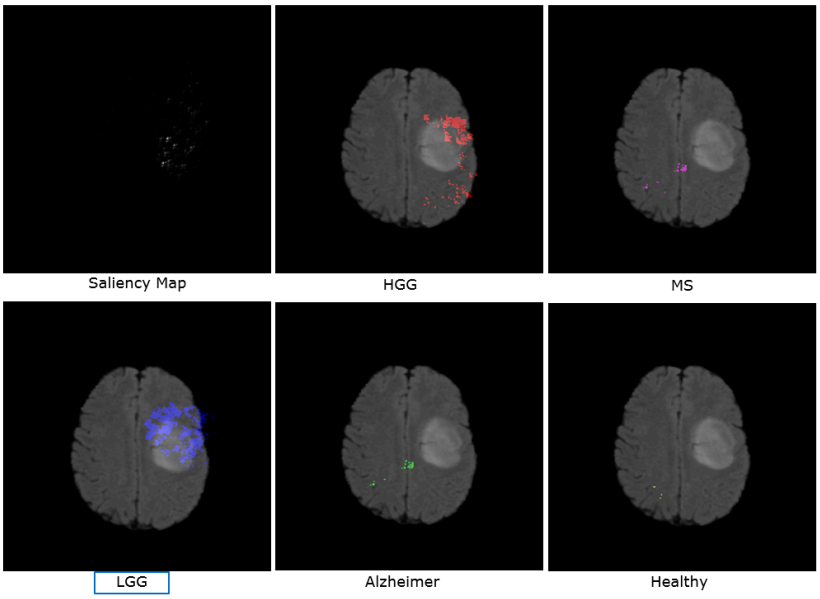


Figure 2.11: Image saliency map shows pixels that are most important for the image being classified as a LGG glioma from ResNet 18 trained with balanced dataset.

2.2.2 Experiments on Patient-wise Batch Normalization

We validated the performance of proposed patient-wise batch normalization on two different implemented generative adversarial algorithms: conditional recurrent-GAN and 3D-GAN. We applied the patient-wise batch normalization in recurrent-GAN for semantic segmentation of MRI cardiac, CT liver and lesions, and microscopic cell images. 3D-GAN framework with patient-wise batch normalization is used for image semantic segmentation of brain tumor. The detail of network architecture is described in Section 2.2.2.

To compare and analysis with other state-of-the-art approaches, we choose four recent public medical imaging challenges in segmentation task: real patient data obtained from the MICCAI 2017, automated Brain MRI segmentation challenge (BraTS-2017), the 2015 microscopic cell segmentation, the ACDC 2017 for simultaneous myocardium and dual cavities segmentation, and 2017 LiTS for semantic segmentation of large liver and small lesions.

Dataset and Preprocessing

LiTS. We applied the LiTS-2017 benchmark [1] that comprised of 130 CT training and 70 test subjects. The examined patients were suffering from different liver cancers. The challenging part is segmentation of very small lesion target on a high unbalanced dataset. Here, pre-processing is carried out in a slice-wise fashion. We applied Hounsfield unit (HU) values, which were windowed in the range of [100, 400] to exclude irrelevant organs and objects. Furthermore, we applied histogram equalization to increase the contrast for better differentiation of abnormal liver tissue.

BraTS. The segmentation of the brain tumour from medical images is highly interesting in surgical planning and treatment monitoring. The goal of segmentation as described by organizer [2, 3, 4, 5] is to delineate different tumour structures such as active tumorous core (TC), enhanced tumorous (ET), and edema or whole

[1] https://competitions.codalab.org/competitions/17094

tumorous (WT) region.

Microscopic Cell. We evaluated the performance of the proposed batch normalization technique on small size microscopic light dataset from human breast carcinoma cells with 1:5 imbalanced class ratio.

ACDC. The ACDC dataset [1] comprised of 150 patients with 3D cine-MR images acquired in a clinical routine. The training database was composed of 100 patients. For all these data, the corresponding manual references were given by a clinical expert. The testing database consisted of 50 patients without manual references.

Network Architecture and Configuration

Conditional recurrent GAN

The RNN-GAN architecture is implemented based on Keras [59] and TensorFlow [60] library. The implemented code is available on my GitHub [2]. All training was conducted on a workstation equipped with double NVIDIA TITAN X GPUs.

The recurrent GAN [61] consists of a generator and a discriminator. The generator is encoder-decoder network that bottleneck is substituted with bidirectional LSTM layer in between. The discriminator is fully convolutional network with bidirectional LSTM layer.

The model is trained for up to 120 epochs with batch size 10, iteration 450 and initial learning rate 0.001 on ACDC dataset. Similarly, in LiTS, we had initial learning rate 0.001, batch size 10, iteration 2750, and 100 epochs where we used all 2D slices from coronal, sagittal, and axial planes with size 256×256. The generator and discriminator for all layers use the *tanh* activation function except the output layer which uses *softmax*. We use categorical cross-entropy as an adversarial loss mixed with categorical accuracy and $\ell 1$. The RMSprop optimizer was used in both the generator and the discriminator. The RMSprop divides the

[1] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html
[2] https://github.com/HPI-DeepLearning/Recurrent-GAN

learning rate by an exponentially decaying average of squared gradients. With this implementation, we are able to produce a cardiac segmentation mask between 500-700 ms per patient on same cardiac phase from ACDC dataset on an axial plane.

The training took eight hours on ACDC for a total of 120 epochs on parallel NVIDIA TITAN X GPUs and with same configuration, it was 12 hours on LiTS dataset.

The proposed approach is trained on 75% training data released by the microscopic cell 2015 and LiTS-2017 benchmarks. We used all provided images from three axes of sagittal, coronal, and axial for training, validation and testing. We trained our system on 75 exams from axial, coronal, and sagittal plane and validated it on the remaining 25 exams for the ACDC dataset.

In both the training and testing phase, the mini-batch consists of 2D images from the same patient, the same acquisition plane and same cardiac phase. We initially normalize the inputs where the stra number is similar to number of classes for semantic segmentation. The mean and variance are computed on a specific patient from the same acquisition plane and from all available images in the same cardiac phase (ED, ES). This normalization helps to restrict the effect of outliers. With batch norm, we normalized the inputs (activation coming from the previous layer) going into each layer using the mean and variance of the activation for the entire mini-batch.

3D-GAN

Our proposed 3D-GAN is implemented based on a Keras library [59] with backend Tensorflow [62] supporting 3D convolutional network and is publicly available [1]. The learning rate was initially set to 0.0001. The Adadelta optimizer is used in both the generator and the discriminator that continues learning even when many updates have been done. The model is trained for up to 200 epochs on BraTS dataset.

Here, the generator network is a modified UNet architecture that we designed two UNet architecture with sharing circumvent bottlenecks and last fully convolutional layer in decoder part. The UNet architecture allows low-level features

[1] https://github.com/HPI-DeepLearning/VoxelGAN

to shortcut across the network. Motivated by previous studies on interpreting encoder-decoder networks [63], that show the bottleneck features carried local features and fully convolutional up-sampling encoder represented global features, we concatenate circumvent bottlenecks and last fully convolutional layer to capture more important features.

Our discriminator is fully convolutional Markovian PatchGAN classifier [64] which only penalizes structure at the scale of image patches. Unlike, the PathGAN discriminator introduced by Isola et al. [64] which classified each N N patch for real or fake, we have achieved better results for task of semantic segmentation in voxel level $1 \times 1 \times d$ we consider N=1 and different d = 64, 32, 16, and 8. We used categorical cross entropy [65] as an adversarial loss with combination of $\ell 1$ loss in generator network.

Evaluation

Figures 2.12 and 2.14 compare the qualitative results from test set when the network was trained with and without patient-wise mini-batch normalization. The patient-wise mini-batch normalization provided normalization for any layer of neural network based on all available 2D images from same patient. Through patient-wise normalization technique, we normalized the actuation's of the previous layer for each patient batch.

Based on qualitative results and Figure 2.12, our network is able to learn from few samples (MDA231) as well as large sample dataset (BraTS2017). We compared quantitative results with the state-of-the-art segmentation method. The quantitative results of individual cell segmentation are detailed in Table 2.7. Obviously, we can see that diversity and the number of images did not have a major effect on the final result.

As shown in Figure 2.13 and Table 2.7 the Gaussian noise negatively influence the segmentation results especially when the trained dataset has few samples. We had same policy for data augmentation on all datasets. We explored during raining the large dataset, when the generator network takes Gaussian noise vector besides medical images, act mostly same as without noise vector and there is minimum differences in the output samples. In contrast, trained network with few samples along with noise vector has negative effect on the final outputs.

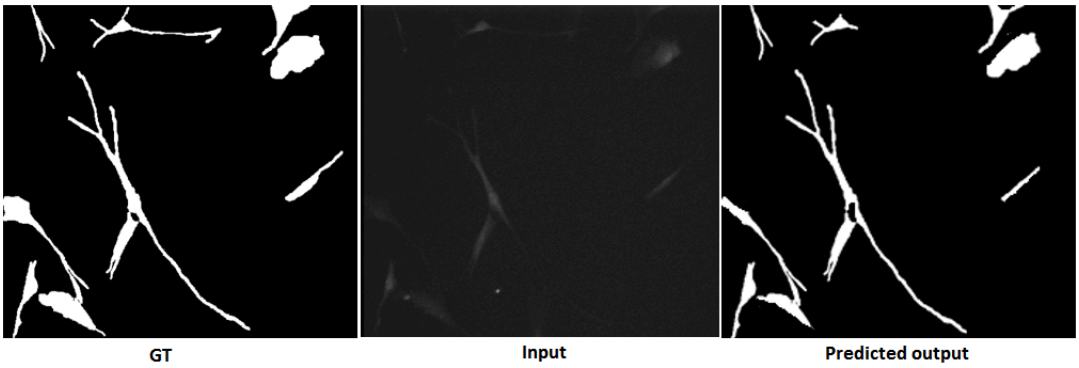


Figure 2.12: Microscopic cell segmentation results obtained by recurrent-GAN network train in patient-wise mini-batch normalization setting and without Gaussian noise.

The results provided in Table 2.11, and Table 2.7 show the improvement of results on both 3D-GAN as well as recurrent-GAN network by patient-wise batch normalization. The qualitative results on Figure 2.15 confirmed the improvement achieved by patient-wise batch normalization for brain tumor segmentation.

Table 2.7: The achieved accuracy for cell segmentation on the MDA231 data, the RNN-GAN (1) shows the results based on patient-wise batch normalization.

Approaches	Dsc	Spec	Sen	FPR	FNR
RNN-GAN (1)	0.93	0.90	0.91	0.10	0.09
RNN-GAN (2)	0.90	0.89	0.91	0.11	0.09
UNet [66]	0.91	-	-	-	
KTH-SE [67]	0.79	-	-	-	
MSER [68]	0.75	-	-	-	
Greedy [69]	0.85	-	-	-	

The qualitative results of liver lesion segmentation are presented in Figure 2.16. Based on Figure 2.16 and Table 2.8, RNN-GAN trained with patient-wise batch normalization is able to detect accurately structure of the lesions. The RNN-GAN architecture trained with patient-wise batch normalization and complementary masks (the results in third rows of Table 2.8) yields better results and trade off between Dice and sensitivity. Dice score is a good measure for class imbalance where indicate the true positive rate by considering false negative and false positive pixels. The effect of class balancing can be seen with comparison of first

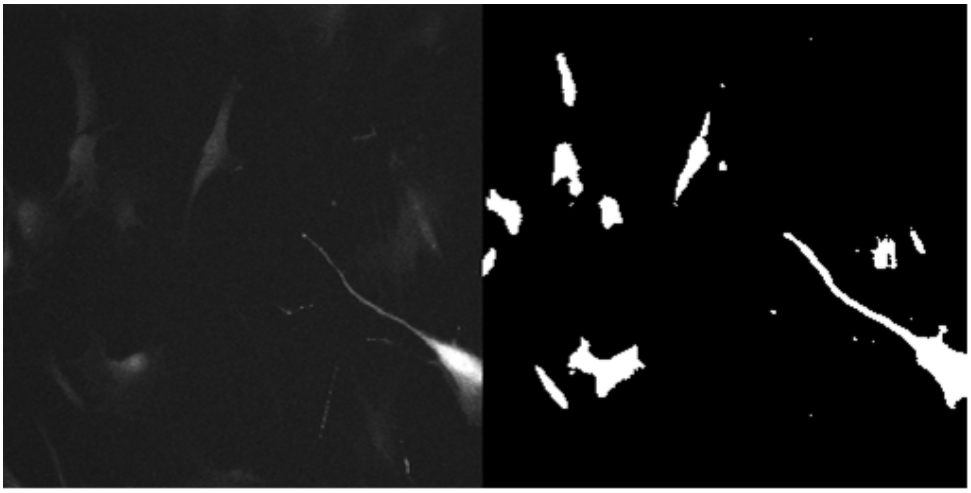


Figure 2.13: Microscopic cell segmentation results obtained by cGAN when the cGAN model trained with additional Gaussian noise as input.

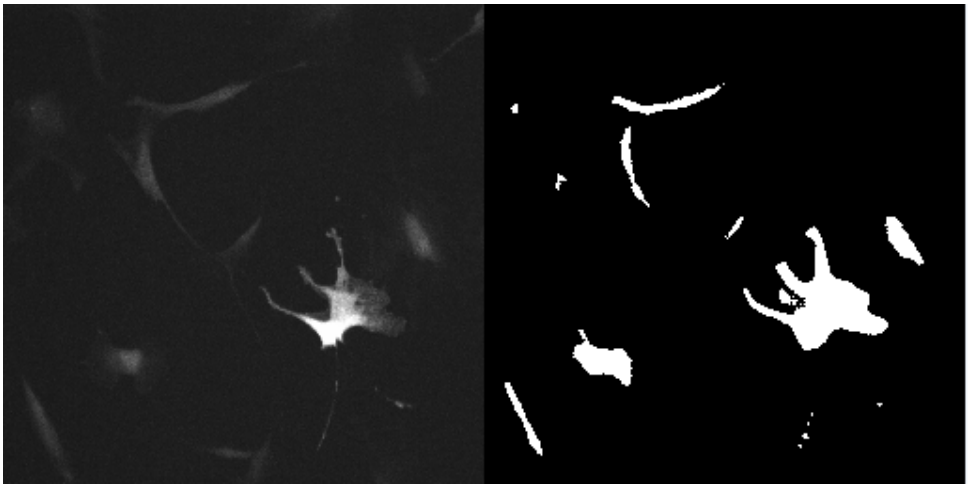


Figure 2.14: Microscopic cell segmentation results obtained by cGAN without patient-wise mini-batch normalization.

and second row of Table 2.8. As we expected the RNN-GAN trained with complementary segmentation labels and binary segmentation masks computed more accurate result with average 2% and 5% improvement respectively in Dice and sensitivity.

We compared predicted results by RNN-GAN at test time with other top-ranked and related approaches on LiTS-2017 in terms of volume overlap error (VOE), relative volume difference (RVD), average symmetric surface dis-

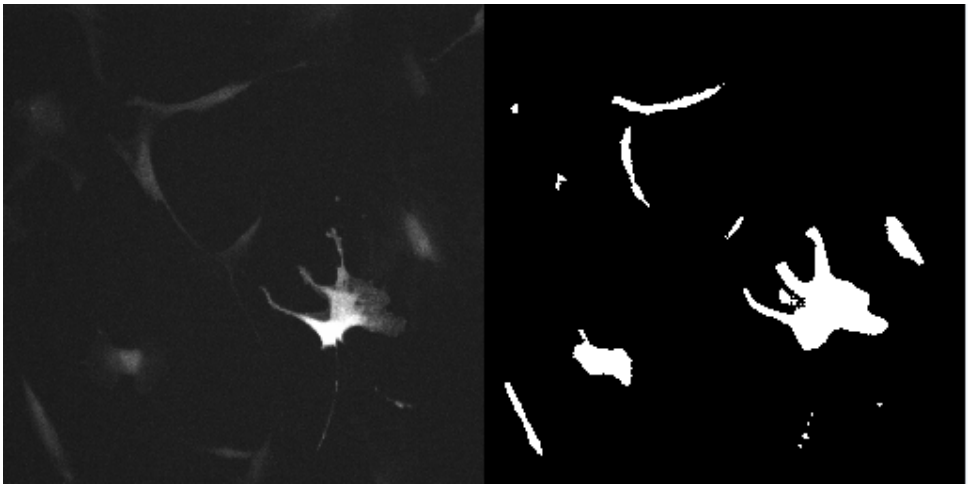


Figure 2.15: Brain tumor segmentation using 3D-GAN framework, the left side images show the segmentation results when the network is trained by conventional batch-normalization. The right side images computed by 3D-GAN considering same configuration, here the network is trained with patient-wise batch-normalization.

tance (ASD), and maximum surface distance or Hausdorff distance (HD), as introduced by challenge organizer. As depicted results in Table 2.8 cascade UNet [70] or ensemble network [71, 72] architectures has achieved better performance compared to trained only with fully convolutional neural network (FCN) [73]. In contrast to prior work such as [70, 71, 72], our proposed method could be generalized to segment the very small lesion and also multiple organs in medical data in different modalities.

The evaluation and comparison performed using the quality metrics introduced by ACDC challenge organizer. In this experiment, semantic segmentation masks are evaluated in a five-fold cross-validation. For each patient, a corresponding images for the End Diastolic (ED) instant and for the End Systolic (ES) instant has provided. As described by ACDC-2017, cardiac regions are defined as right-ventricle region labeled 1, 2 and 3 representing respectively myocardium and left ventricles.

As shown in Table 2.9, our method outperforms other top-ranked approaches from the ACDC benchmark when the RNN-GAN is trained with patient-wise batches and complementary masks. Based on Table 2.9, in Dice coefficient, our method achieved slightly better than the Wolterink et al. [74] on ACDC challenge

Table 2.8: Quantitative segmentation results for liver and lesions segmentation on the LiTS-2017 dataset. The first and second rows show achieved accuracy for the task of liver lesions segmentation when our network is trained with patient-wise batch normalization RNN-GAN (1), with conventional batch normalization RNN-GAN (2). RNN-GAN* shows results with complementary segmentation masks and patient-wise which we briefly explain it as future direction. The columns Dice 1 and Dice 2 show the segmentation results for liver and lesion respectively.

Architecture	Dice 1	Dice 2	Sen 1	Sen 2	VOE	RVD	ASD	HD
RNN-GAN (1)	0.93	0.80	0.81	0.73	20	-2	9.7	52.3
RNN-GAN (2)	0.89	0.78	0.69	0.63	22	3	11.4	58.1
RNN-GAN *	0.94	0.83	0.74	0.74	14	-6	6.4	40.1
cGAN	0.88	0.76	0.57	0.55	21	-1	10.8	87.1
UNet [70]	0.82	0.72	-	-	22	-3	9.5	165.7
ResNet+Fusion [71]	0.86	-	-	-	16	-6	5.3	48.3
H-Dense+ UNet [72]	0.89	-	-	-	39	7.8	1.1	7.0
FCN [73]	-	-	-	35	12	1.0	7.0	

in left ventricle and myocardium segmentation. However, Rohe et al. [75] achieved outstanding performance for right ventricle segmentation since they applied the multi-atlas registration and segmentation at the same time. Zotti et al. [76] achieved competitive results based on GridNet for left ventricle segmentation with overall Dice 0.96 and 0.94 for right ventricle segmentation.

Based on Table 2.9 and Table 2.10, the right ventricle is a difficult organ for all the participants mainly because of its complicated shape, the partial volume effect close to the free wall, and intensity of homogeneity. Our achieved accuracy in term of Hausdorff distance, in average is $1.2 \pm 0.2mm$ lower than other participants. This is a strong indicator for precision of boundary that RNN-GAN architecture substituted with bidirectional LSTM units is suitable solution for capturing the temporal consistency between slices. Compared to cGAN (Table 2.9 and Table 2.10) RNN-GAN (1) provides better results when the network is trained with patient-wise batch normalization setting and even sensitivity and precision.

Compared to the expert annotated file on the original ED phase instants, individual Dice scores of 0.968 for the left ventricle (LV), 0.933 for the myocardium (MYO), and 0.940 for the right ventricle (RV) (see Table 2.9) were

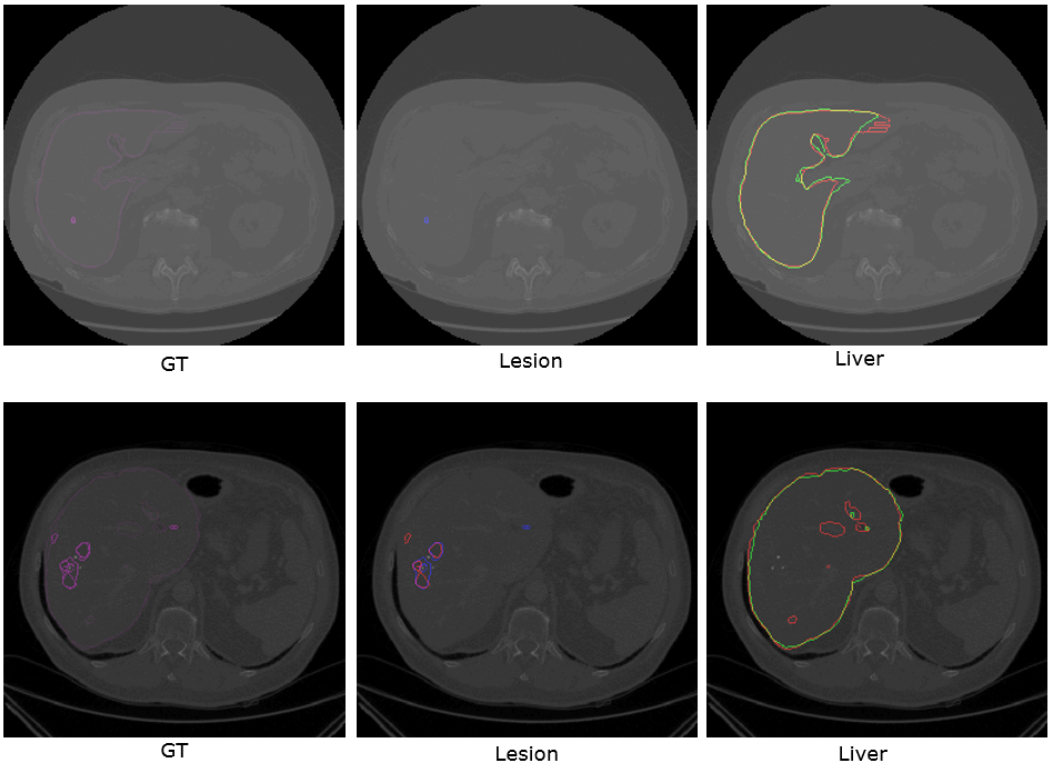


Figure 2.16: LiTS-2017 test results for simultaneous liver and lesion segmentation using RNN-GAN. The first row shows the results when the RNN-GAN trained with conventional batch normalization way while second row presents result with proposed patient-wise batch normalization. The first column is ground truth annotated by medical expert, blue and purple color in second column code the ground truth and predicted lesion border by RNN-GAN. Yellow and red color boundaries in third column show the ground truth and predicted liver region by our proposed method.

achieved in test time on 25 patients. Qualitatively, the RNN-GAN segmentation results are promising (see Figure 2.17) where we can see robust and smooth boundaries for all substructures.

As depicted on Figure 2.17 and Table 2.9 right ventricle is complex organ to segment. The most failure happened in systolic phase. Based on Figure 2.17 the achieved accuracy in the test time on ACDC benchmark, we observed that the average results in diastolic phase (first and second row) are better than the average results on systolic phase (third and fourth row). The evaluated quantitative results trained by RNN-GAN*, with patient-wise and complementary labels in term of Hausdorff distance and Dice are shown in Figure 2.18. As expected, the

Table 2.9: Comparison of the achieved accuracy in term of Dice metric on ACDC benchmark with related approaches and top-ranked methods. RNN-GAN (1), RNN-GAN (2), RNN-GAN* trained with, without patient-wise batch normalization, and with both the ground truth and complementary masks respectively.

Methods	Phases	Left Ventricle	Right Ventricle	Myocardium
RNN-GAN (1)	ED	0.961	0.940	0.933
	ES	0.944	0.911	0.925
RNN-GAN (2)	ED	0.962	0.934	0.926
	ES	0.928	0.90	0.917
RNN-GAN*	ED	0.968	0.940	0.933
	ES	0.951	0.919	0.925
cGAN	ED	0.934	0.906	0.899
	ES	0.918	0.874	0.870
Isensee et al. [77]	ED	0.955	0.925	0.865
	ES	0.905	0.834	0.882
Wolterink et al. [74]	ED	0.96	0.92	0.86
	ES	0.91	0.84	0.88
Rohe et al. [75]	ED	0.94	0.96	0.90
	ES	0.92	0.95	0.90
Zotti et al. [76]	ED	0.96	0.94	0.89
	ES	0.94	0.87	0.90
U-Net [66]	ED	0.96	0.88	0.78
	ES	0.92	0.79	0.76
ConvDeconv	ED	0.92	0.82	0.76
	ES	0.87	0.64	0.81
Poudel et al. [78]		0.90	-	-

achieved Dice score on left ventricle (median of 6.82/8.02 for the ED/ES frames) tend to be lower than for the two other regions of interest with myocardium at 8.08/8.69 and right ventricle at 8.95/12.07 for ED/ES.

We evaluated the performance of proposed patient-wise batch normalization with 3D-GAN framework. We tested 3D-GAN with and without patient-wise batch normalization on BraTS2017 dataset for semantic segmentation of brain tumors.Table 2.11 shows the reported results evaluated by online platform of

Table 2.10: Comparison of achieved accuracy in term of Hausdorff distance on ACDC benchmark with top-ranked participant approaches and related work.

Methods	Phases	Left Ventricle	Right Ventricle	Myocardium
RNN-GAN (1)	ED	7.12	9.21	8.72
	ES	8.11	12.72	8.639
RNN-GAN (2)	ED	8.62	12.16	9.04
	ES	9.44	13.2	9.50
RNN-GAN*	ED	6.82	8.95	8.08
	ES	8.02	12.17	8.69
cGAN	ED	8.62	12.16	9.04
	ES	9.44	13.2	9.50
Isensee et al. [77]	ED	7.38	10.12	8.72
	ES	6.90	12.14	8.67
Wolterink et al. [74]	ED	7.47	11.87	11.12
	ES	9.6	13.39	10.06
Rohe et al. [75]	ED	7.04	14.04	11.50
	ES	10.92	15.92	13.03
Zotti et al. [76]	ED	5.96	13.48	8.68
	ES	6.57	16.66	8.99
U-Net [66]	ED	6.17	20.51	15.25
	ES	8.29	21.20	17.92
ConvDeconv	ED	8.77	22.59	13.92
	ES	10.34	28.45	11.64

BraTS challenges. Figure 2.19 shows qualitative results trained by patient-wise batch normalization 3D-GAN.

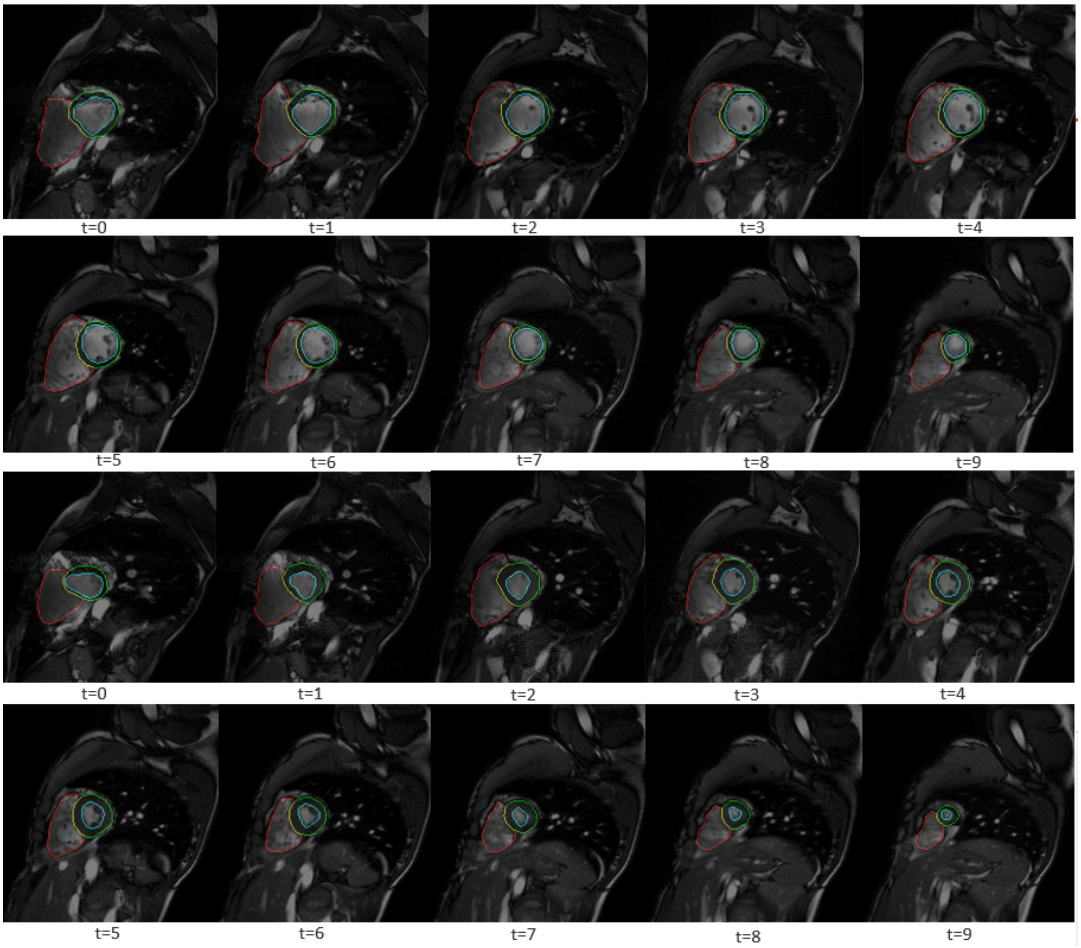
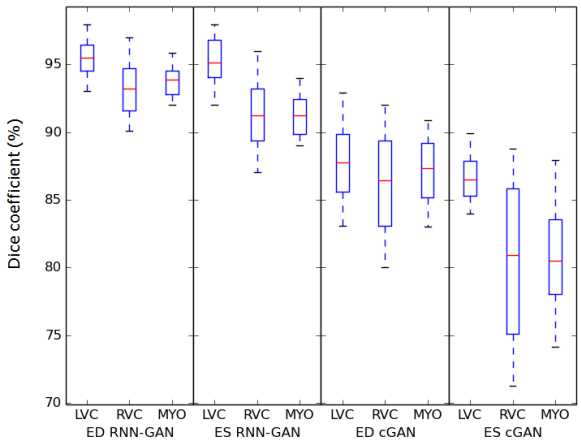


Figure 2.17: The cardiac segmentation results at test time by RNN-GAN from ACDC 2017 benchmark on Patient084. The red, green, and blue contour present respectively right ventricle, myocardium, and left ventricle region. The top two rows show the diastolic phase from different slices from t=0 till t=9 circle. Respectively the third and fourth rows present systolic cardiac phase from t=0 till t=9 circle.

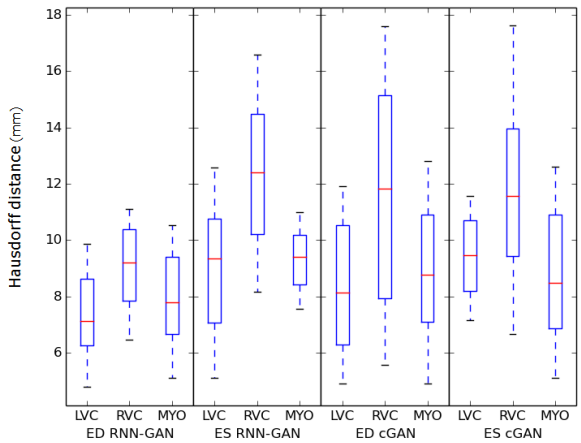
2.3 Related Works

This section briefs the previous studies carried out on synthetic sampling technique using generative adversarial network, batch normalization method and biased with complementary labels mostly in recent years.

Synthesis Minority Over Sampling using Generative Adversarial Net-



(a)



(b)

Figure 2.18: The ACDC 2017 challenge results using RNN-GAN* and cGAN architecture. The left figure shows Dice coefficient in two cardiac phase as follows the right sub figure presents Hausdorff distance. The y-axis shows the Dice metrics and x-axis shows segmentation performance based on cGAN and RNN-GAN* in ED and ES cardiac phase. In each sub figure, the mean is presented in red. The ACDC 2017 challenge results using RNN-GAN* and cGAN architecture. The sub figure (b) y-axis codes the Hausdorff distance in mm and x-axis presents segmentation performance based on cGAN and RNN-GAN* in ED and ES cardiac phase.

Table 2.11: Comparison results of the achieved accuracy for semantic segmentation by 3D-GAN when the trained model with patient-wise batch normalization 3D-GAN (1), and with conventional batch normalization 3D-GAN (2). The reported results evaluated by online platform of BraTS challenge in terms of Dice, sensitivity (Sen), specificity (Spec), and Hausdorff distance (Hdff). WT, ET, and CT are abbreviation of whole tumor, enhanced tumor, and core of tumor regions respectively.

	Dice			Hdff			Sen			Spec		
Methods	WT	ET	CT	WT	ET	CT	WT	ET	CT	WT	ET	CT
3D-GAN(1)	0.83	0.61	0.68	6.81	8.3	10.72	0.80	0.73	0.63	0.99	0.99	0.99
3D-GAN(2)	0.81	0.61	0.64	7.30	9.2	12.04	0.75	0.61	0.55	0.99	0.99	0.99

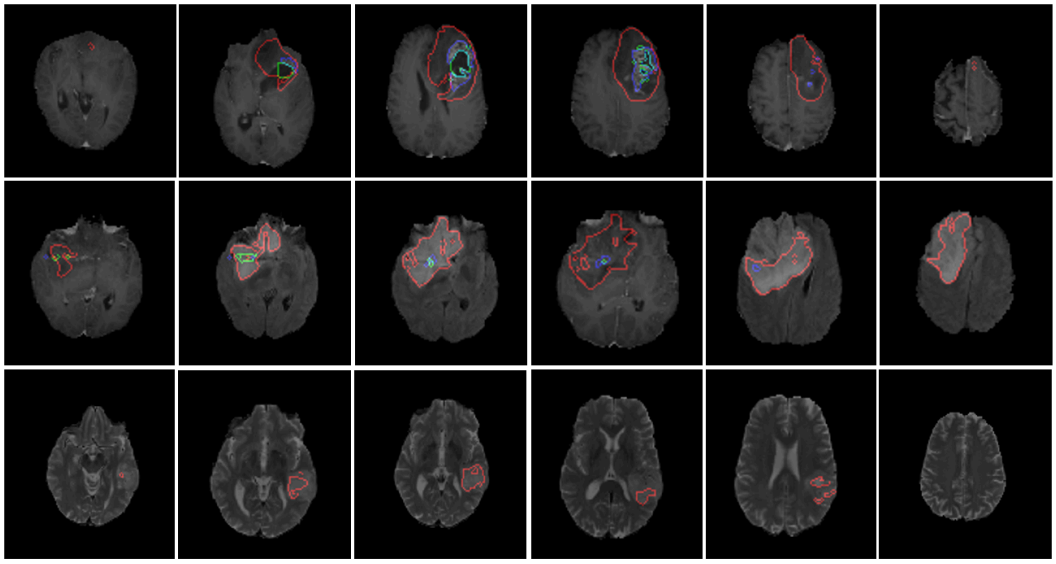


Figure 2.19: BraTS-2017 test results for semantic segmentation trained by 3D-GAN with patient-wise batch normalization. Red, blue, and green colors represent segmentation borders for whole tumor, enhanced tumor, and tumor core respectively.

works

The synthesize of realistically looking medical images opens up many new opportunities to tackle well-known deep learning problems such as class imbalance, data augmentation, or the lack of labeled data. Kazeminia et al [79] reviews recent GAN networks that have been used in the literature of medical imaging. The recent GAN architectures applied for medical image analysis have been used

in two settings: unsupervised which synthesize certain types of medical images from the noise alone, or in a conditional way, the prior knowledge such as meta-data or even image data used as an additional noise to generative network. Initial results have shown the success of DCGAN architecture can be used to synthesize realistically looking patches of retinal images [80], CT liver lesion [81], abdomen lung [82]. However, these approaches applied the DCGAN architecture to synthesize patches mostly at a resolution of 1616 pixels and 64 64 pixels, respectively, while we trained DCGAN to mimic the MRI distribution in an unsupervised manner.

Conditional GAN widely used on medical images recently for synthesis [31, 80], segmentation [83], reconstruction [32], detection [84], registration [85], and classification [86]. The radiation exposure by CT images put the patient at risk of cell damage and cancer which motivated to synthesize CT images from MRI. Nie et al. [31] synthesize CT images from corresponding MR images with the help of a cascade of fully convolutional networks which they train with a normal reconstruction loss, an image gradient loss and additionally with an adversarial network in order to improve realism of the synthetic CT images.

In this chapter, we synthesize different MRI modalities to generate minority class samples and improve classification and detection results using deep multi-modal network [87]. We applied deep generative adversarial network in unsupervised setting and generated samples for minority classes. In a different experiment, we implemented conditional recurrent fully convolutional networks which substituted by bidirectional LSTM in bottleneck to generate one or two missing MRI modalities. The discriminator is recurrent fully convolutional classifier [61]. Then, the generated samples are used to balance data distribution in multi-modality setting and show better performance for brain diseases diagnosis.

Batch Normalization Techniques

Batch normalization is a method that normalizes the inputs of each layer in order to fight the internal covariate shift problem. Dong et al. [37] introduced batch-wise mining to tackle with imbalanced of minority class incremental rectification using a deep convolutional neural network. Recently, Ioffe [47] introduced a new Batch Renormalization since the conventional batch normalization is not well

suited to small training set consists of different samples, which is one nature problem by imbalanced samples. Batch Renormalization [47] replaces batchnorm to ensure that the outputs computed by the model are dependent only on the individual examples and not the entire dataset, during both training and testing. In this chapter, we reformulated batch normalization by stratified sampling since most of clinical dataset and medical imaging dataset are small and more homogeneous. Our learning algorithm produces inductive bias (learning bias) towards the frequent (majority) and a minority of training data.

Learning with complementary labels

We presented some preliminary results based on the complementary labels that we plan to extend this chapter in this direction. The complementary labels in context of machine learning [88] has been used by assuming the transition probabilities are identical with modifying traditional one-versus-all and pairwise-comparison losses for multi-class classification. Ishida et. al. [88] theoretically prove that unbiased estimator to the classification risk can be obtained by complementary labels. Yu et al. [89] study learning from both complementary labels and ordinary labels can provide a useful application for multi-class classification task. Inspired by recent success [88, 89], we train the proposed RNN-GAN with both complementary labels and ordinary labels for the task of semantic segmentation to skew the bias from majority pixels (see Section 2.2).

2.4 Extensions and Summary

Future Direction: Biased with Complementary Labels

As future research in data-level, we study internal bias correction by inverse class frequency labels. The complementary labels developed motivated by surrogates of training data with true label. The complementary labels specify a class that feature does not belong to. Given the complementary labels to the less represented samples shown better performance by assuming the transition probabilities are identical with modifying traditional one-versus-all and pairwise-comparison

losses for multi-class classification [88]. We plan to explore the advantage of training network simultaneously with true labels and the inverse of class frequency, named complementary label. For example in task of semantic segmentation in addition to ground truth segmentation masks (ordinary masks), complementary masks for majority class is provided. Here, the majority and minority pixels value are changed to skew bias from majority pixels where the negative label for the major class and a positive label for the $c-1$ class. We presented our preliminary results in Table 2.9 and Table 2.8 that represent better performance for multi label (semantic) segmentation. Therefore, we see potential of research in this direction and plan to apply these approach on more dataset for different tasks and provide theoretical guarantee for these approach.

Summary

In this chapter, we proposed two new approaches to correct internal bias and mitigate imbalanced class problem for brain disease diagnosis and semantic segmentation. We demonstrated minority class oversampling using deep generative adversarial network in unsupervised and supervised manner can alleviate bias from majority class. We balance training set distribution by selecting synthesized samples with minimum FID, and tested for brain disease diagnosis. In different experiment, we synthesized missing image modalities and observed better performance for classification using trained model with multi-modalities image.

In this chapter, we demonstrated a patient-wise batch normalization to modified internal bias. Each mini-batch sample are constructed within homogeneous among heterogeneous. We validate our technique on two different generative models; recurrent conditional GAN and 3D-GAN for semantic segmentation. The experiments on different medical imaging benchmarks demonstrated the generalization ability of our approach for segmentation of body organ and tumor regions.

Chapter 3

Instance Weighting for Mitigating Imbalanced Data

In this chapter, we address the problem of imbalanced data by modifying existing learners to alleviate their bias towards majority groups. We adjust the learning algorithm to incorporate varying penalty for each considered group of examples. In this chapter, we introduce two different policies for weighing loss function. As a first approach, we assign a higher cost to the less represented set of objects and boost its importance during the learning process. In the second approach, a loss function is weighted to reduce the effect of majority class frequency.

The proposed cost sensitive weighting losses are tested with two different generative adversarial frameworks on different tasks; semantic segmentation and diseases classification. Throughout this work, we have shown that the weighted loss function can mitigate imbalanced training data and improve segmentation results for highly imbalanced brain tumor segmentation, cardiac MRI segmentation, and less imbalanced microscopic cell segmentation. At the end of the chapter, we discuss insights gained from segmentation and diseases prediction, explain variety of applications and extensions, and suggest research direction.

3.1 Background

Cost-sensitive learning approaches assume the different cost based on misclassification. A cost matrix determines by the penalty of classifying samples from one class as another. Assume $C(i, j)$ denote the cost of predicting an instance from class i as class j. Let $C(+, -)$ denote the cost of predicting a positive instance as a negative instance, and $C(-, +)$ is the cost of the contrary case. In the imbalanced dataset, the importance of positive instances is higher than the negative instances. The cost of misclassified a positive instance outweighs the cost of misclassified a negative one (i.e. $C(+, -) \succ C(-, +)$); predicting a correct classification present zero penalty (i.e. $C(-, -) = C(+, +) = 0$).

The cost-sensitive approaches try to minimize the misclassification cost and the number of high-cost errors while training a model. The cost-sensitive methods can be categorized into two categories: (1) learning a specific classifier and (2) instance weighting.

The learning specific classifier tries to modify the training data distribution to minimize error regarding misclassification costs such as Dice coefficient loss [87, 90, 91], and asymmetric similarity loss [92]. However, their costs are chosen in multiple runs of the network and remain fixed during the learning process at each time step. In contrast, we modify the cost function by weighting instance from minority class [93] and weighting all instances to balance data distribution [94]. Zadrozny et al. [95] present empirical results and explain the decision theorem and translation theorem guarantees the performance of weighting classification regarding minority class instances.

3.2 Cost Instance Weighting

We mitigate the imbalanced class problem by introducing two new instance weighting losses. The proposed losses are tested based on: (1) a deep voxel-GAN framework [93] for semantic segmentation which the cost function is weighted to reduce the effect of majority class by an average of minority classes, (2) a

recurrent-GAN [61] for learning semantic segmentation and disease classification that balance feature space on by weighing both minority and majority classes. In the following, we describe the detail of each loss on two different network architecture.

3.2.1 voxel-GAN

We proposed voxel-GAN consists of a segmentor network that takes 3D multimodal MR or CT images x and Gaussian vector z. It outputs a 3D semantic segmentation; $S : \{x, z\} \rightarrow \{y_{seg}\}$. The discriminator takes the segmentor output $S(x, z)$ and the ground truth annotated by an expert y_{seg} and classifies a confidence value $D(x)$ of whether a 3D object input x is synthetic or real. The training procedure is similar to a two-player mini-max game, as shown in Eq.(3.1).

$$\begin{aligned}\mathcal{L}_{adv} \leftarrow \min_{S} \max_{D} V(D, S) = E_{x,y_{seg}}[logD(x, y_{seg})]+ \\ E_{x,z}[log(1 - D(x, S(x, z)))]\end{aligned} \tag{3.1}$$

Here, the segmentor loss is weighted the same as Eq.(3.2) to reduce the effect of majority class voxel frequencies for the dataset.

$$w_i = \begin{cases} avg\{f_i\}\{0 < i < c\}/f_{max}, & \text{if } i \text{ is max frequency} \\ 1, & \text{otherwise} \end{cases} \tag{3.2}$$

The segmentor loss Eq.(3.3) is mixed with $\ell 1$ term to minimize the absolute difference between the predicted value and the existing largest value. Hence, the $\ell 1$ objective function takes into account CNNs feature differences between the predicted segmentation and the ground truth segmentation and resulting in fewer noises and smoother boundaries.

$$\mathcal{L}_{L1}(S) = E_{x,z} \parallel y_{seg} - S((x * w_i), z) \parallel \tag{3.3}$$

$$\mathcal{L}_{seg}(D, S) = \mathcal{L}_{adv}(D, S) + \mathcal{L}_{L1}(S) \tag{3.4}$$

The final objective function for semantic segmentation of brain tumors $\mathcal{L}_{seg}$ calculated by adversarial loss and additional weighted $\ell 1$ loss (see Eq.3.4).

Segmentor Architecture

As shown in Fig. 3.1, the segmentor architecture is two, 3D fully convolutional encoder-decoder network that predicts a label for each voxel. The first encoder takes 128 × 128 × 128 of multi-modal MRI or CT images at the same time as different channel input. Last decoder outputs 3D images with size 128 × 128 × 128. Similar to UNet [66], we added the skip connections between each layer i and layer $n-i$, where n is the total number of layers in each encoder and decoder part. Each skip connection concatenates all channels at layer i with those at layer $n-i$. The bottleneck features are concatenated with last convolutional decoder to capture better feature representation.

Discriminator Architecture

As depicted in Fig. 3.1, the discriminator is 3D fully convolutional encoder network which classifies whether a predicted voxel label belongs to the right class. More specifically, the discriminator is trained to minimize the average negative cross-entropy between predicted and the correct labels.

Then, the segmentor and the discriminator networks are trained through back propagation corresponding to a two-player mini-max game. We use categorical cross entropy [65] as an adversarial loss. As mentioned before, we weighted loss to only attenuate healthy voxel impact in training and testing time.

3.2.2 recurrent-GAN

Similar conditional GAN [11]; the proposed recurrent-GAN, a generative model learns mapping from a given sequence of 2D multimodal MR images x_i to a sequence semantic segmentation $y_{i_{seg}}$ and classification y_{cls}; $G : \{x_i, z\} \rightarrow \{y_{i_{seg}}, y_{cls}\}$ (e.g. i refers to 2D slice index between 1 and 155 from a total 155 slices acquired from each patient). The training procedure for the segmentation task is similar to two-player mini-max game, as shown in Eq.(3.5). While the generative model generated segmentation pixels label, the discriminator classifies whether the predicted pixel output by the generator is similar to the ground truth annotated by

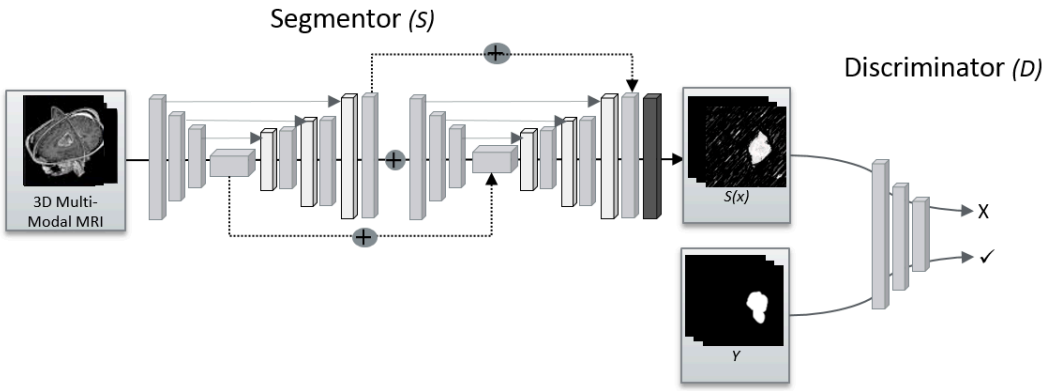


Figure 3.1: The proposed voxel-GAN consists of a segmentor network S and a discriminative network D. S takes 3D multi modal images as a condition and generates the 3D semantic segmentation as outputs, D determines whether those outputs are real or fake. We use modified 3D hourglass as a segmentor network in order to capture local and global features extracted in bottleneck and last convolutional decoder. Here, D is 3D fully convolutional encoder.

a medical expert or synthetic. The adversarial loss is mixed with two additional loss to attenuate the imbalanced data impact.

$$\begin{aligned}\mathcal{L}_{adv} \leftarrow \min_G \max_D V(G,D) = E_{x,y_{seg}}[logD(x,y_{seg})]+ \\ E_{x,z}[log(1-D(x,G(x,z)))]\end{aligned} \tag{3.5}$$

An imbalanced class in a medical dataset where the non-healthy class cannot train as well as a healthy class might dominate the gradient direction. Regarding to mitigate class-imbalanced impact, we mixed adversarial loss (Eq. 3.5) with selective weighted (Eq.3.6) categorical cross-entropy loss $\mathcal{L}_H(G)$ for semantic segmentation, and with selective weighted (Eq.3.6) $\ell 1$ loss (Eq.3.8) for classification of diseases.

$$w_c = \sqrt{\frac{\mid C_c \mid}{f_c + N}} \tag{3.6}$$

Where the weight for each class c is calculated on the ratio of the cordiality among N classes on entire training dataset (e.g., mostly healthy classes) by the frequency of samples with class, c appears in the dataset. Since we intense frequency differences, the square root is applied to prevent huge weights. This

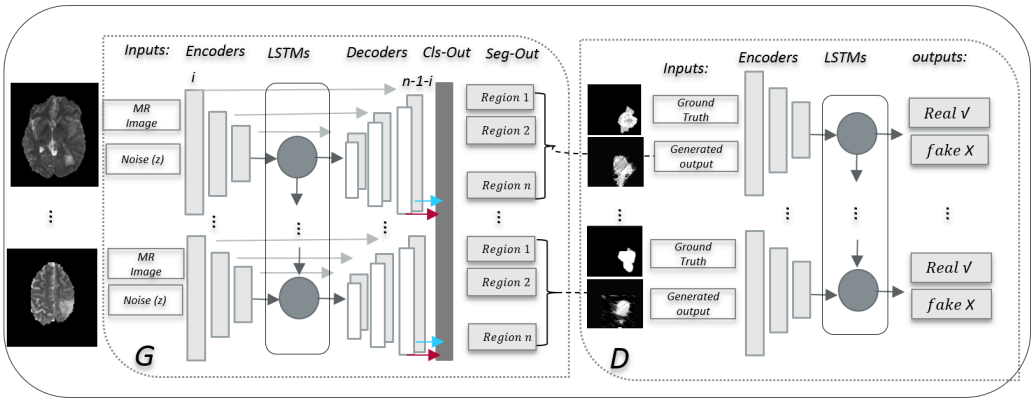


Figure 3.2: Our proposed architecture for learning semantic segmentation and diseases prediction. We design a set of auto-encoders combined with a LSTM unit in a circumvent bottleneck as the generator network with skip connections between each layer i and the corresponding layer n-1-i (mostly like UNet architecture). The discriminator is fully convolutional network substituted with LSTM unit. Both networks are trained together in an adversarial way with selective weighted categorical cross entropy loss for semantic segmentation and selective weighted L1 for diseases prediction.

implies that larger classes in the training set have a weight smaller than 1, and the weights of the smallest classes are the highest defined by Eq.(3.6).

The final loss for semantic segmentation task is calculated through:

$$\mathcal{L}_{seg}(D,G) = \mathcal{L}_{adv}(D,G) + \mathcal{L}_{H}(G * w_c) \tag{3.7}$$

As shown in Fig.3.2, the concatenated depth features from last decoder layer with skip connection of encoder part passed into a couple of dense layers and map to the disease class. The objective function for class prediction is $\ell 1$ to minimize the absolute difference between the predicted value and the existing largest value Eq.(3.8)

$$\mathcal{L}_{cls}(G) = E_x \parallel y_{cls} - \sum_{i=1} G(x_i * w_c) \parallel \tag{3.8}$$

Where i indicates to 2D slice index from the same patient (e.g., in Brain dataset i is between 1 and 155 from a total of 155 slices acquired from each patient).

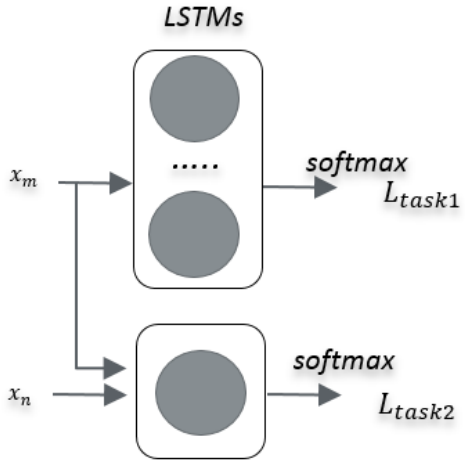


Figure 3.3: learning multi-task using recurrent GAN.

In this work, similar to the work of Isola et al. [64], we used Gaussian noise z in the generator alongside the input data x. As discussed by Isola et al. [64], in training procedure of conditional generative model from conditional distribution $P(y|x)$, that would be better to model produces more than one sample y, from each input x. When the generator G, takes plus input image x, random vector z, then $G(x, z)$ can generate as many different values for each x as there are values of z. Specially for medical image segmentation, the diversity of image acquisition methods (e.g., MRI, fMRI, CT, ultrasound), regarding their settings (e.g., echo time, repetition time), geometry (2D vs. 3D), and differences in hardware (e.g., field strength, gradient performance) can result in variations in the appearance of body organs and tumour shapes [96], thus learning random vector z with input image x makes network robust against noise and act better in the output samples. This has been confirmed by our experimental results using datasets having an extensive range of variation.

As depicted in Figure 3.3, we address learning multi-task by assigning different losses for each task. The final objective function for simultaneous semantic segmentation and classification is similar:

$$\mathcal{L} = \mathcal{L}_{seg}(D, G) + \mathcal{L}_{cls}(G) \tag{3.9}$$

Our proposed architecture consists of a generator network, G, in the left side followed by a discriminator network, D, in the right side of the figure.

Generator Network

The recurrent generator takes a random vector z together with the sequence of MR images. We design a set of auto-encoders combined with an LSTM $n \times n$ unit in a circumvent bottleneck as the generator network.

Similar UNet [66], we keep skipping connections between each layer i and the corresponding layer $n-1-i$, where n represents the total number of layers. Each skip connection concatenates all channels at layer i with those at layer $n-1-i$. Feature maps from the convolution part in the down-sampling step are fed to the up-convolution part in the up-sampling level.

The generator is trained on a sequence of 2D images from the same patient on different multi sites multi-modalities. At the end of each sequence, the output features from decoder, fed to the fully connected layer and map to one of the patient diseases.

Discriminator Network

The discriminator network is a classifier and has a similar structure as an encoder and bidirectional LSTM. Hierarchical features are extracted from the fully convolutional encoder of the discriminator and used to classify between the generator segmentation output and ground truth. Especially, the discriminator is trained to minimize the average negative cross-entropy between predicted and the actual label of an image in pixel level at each time step.

Then, two models are trained through back propagation corresponding to a two-player mini-max game. We use binary cross-entropy [65] as an adversarial loss, a categorical cross-entropy as additional loss for segmentation, and $\ell 1$ loss for classification. In this work, the recurrent architecture selected for both discriminator and generator is a bidirectional LSTM [97].

3.3 Experiments

We validated the performance of proposed weighting losses on three recent medical imaging challenges: real patient data obtained from the MICCAI 2018, brain diseases diagnosis and tumor segmentation (BraTS) [2, 3, 4, 5], cardiac diseases classification and dual cavities and myocardium segmentation (ACDC-2017) [1] dataset, and CT brain lesion segmentation challenge (ISLES-2018) [98, 99].

3.3.1 Datasets and pre-processing

The BraTS 2018 benchmark [2, 3, 4, 5] prepared the data in two different brain diseases; high and low-grade glioma (HGG/LGG) brain tumor (s). The BraTS 2018 released 1156 magnetic resonance images in four modalities T2, Flair, T1, and T1c from 289 patients. All training data annotated and semantically segmented by the medical expert(s) with four segmentation labels, namely non-tumor, necrosis or tumorous core, edema or whole tumor, and enhanced tumor as shown in Figure (3.5). The purpose of this challenge is to segment the complex and heterogeneously located brain tumors automatically and classify brain disease.

The ACDC [1] comprised by 150 patients with 3D cine-MR images from five subgroups of healthy subjects, patients with abnormal right ventricle (RV), hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), and myocardial infarction (MINF). The training dataset released with ground truth file manually segmented by the medical expert(s) for 100 samples (equally for each class 20, 3D cine-MR images) with four segmentation labels of background, right ventricle, myocardium, and left ventricle. The goal of the challenge is to segment the three adjoint regions with a considerable inter-subject variation.

We applied ISLES2018 contains 94 computed tomography (CT) and MRI training data in six modalities of CT, 4DPWI, CBF, CBV, MTT, Tmax, and the annotated ground truth file. The examined patients were suffering from different brain cancers. The challenging part is binary segmentation of unbalance labels. Here, pre-processing is carried out in a slice-wise fashion. We applied the Hounsfield unit (HU) values, which were windowed in the range of [30, 100] to

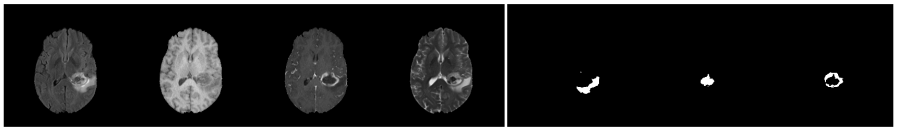


Figure 3.4: The axial view of low grade glioma brain image from BraTS 2017. The fifth, sixth, and seventh column show the whole tumor, core of tumor and enhanced tumorous region.

get soft tissues and contrast. Furthermore, we applied histogram equalization to increase the contrast for better differentiation of abnormal lesion tissue.

We added data augmentation to each dataset such as randomly cropped, resizing, scaling, rotation between -10 and 10 degrees, and Gaussian noise applied on training and testing time for both datasets.

3.3.2 Implementation

This section provides more details in configuration and network architecture by voxel-GAN and recurrent-GAN.

Configuration of voxel-GAN

Our proposed voxel-GAN is implemented based on a Keras library [59] with backend Tensorflow [62] supporting 3D convolutional network and is publicly available [1]. All training and experiments were conducted on a workstation equipped with a multiple GPUs. The learning rate was initially set to 0.0002. The Adadelta optimizer was used in both the segmentor and the discriminator that continues learning even when many updates have been done. The model is trained for up to 200 epochs on each dataset separately.

Network Architecture of voxel-GAN

In the voxel-GAN, a segmentor network is a modified UNet architecture that we designed two UNet architecture with sharing circumvent bottlenecks and last fully convolutional layer in decoder part. The UNet architecture allows low-level features to shortcut across the network. Motivated by previous studies on interpreting encoder-decoder networks [63], that show the bottleneck features carried

[1] https://github.com/HPI-DeepLearning/VoxelGAN

Table 3.1: Comparison of achieved average Dice coefficient per class (per voxel) by different depth and fixed receptive field in discriminator, evaluated on BraTS 2017 (on local validation set).

Methods	WT	ET	CT
$1 \times 1 \times 1$	0.84	0.64	0.79
$1 \times 1 \times 8$	0.75	0.58	0.59
$1 \times 1 \times 16$	0.78	0.68	0.68
$1 \times 1 \times 32$	0.81	0.70	0.86
$1 \times 1 \times 64$	0.69	0.58	0.58
$1 \times 1 \times 96$	0.67	0.57	0.52

local features and fully convolutional up-sampling encoder represented global features, we concatenate circumvent bottlenecks and last fully convolutional layer to capture more important features.

The segmentor takes four available MR image modalities (T1, T2, T1c, and Flair) on BraTS 2017 which each MRI volume size $240 \times 240 \times 155$. We center cropped each subject into a sub volume of $128 \times 128 \times 128$, to remove the border black regions while still keep the entire brain regions.

Our discriminator is fully convolutional Markovian PatchGAN classifier [64] which only penalizes structure at the scale of image patches. Unlike, the PatchGAN discriminator introduced by Isola et al. [64] which classified each $N \times N$ patch for real or fake, we have achieved better results for task of semantic segmentation in voxel level $1 \times 1 \times d$ we consider different $d = 96$, 64, 32, 16, and 8. We used categorical cross entropy [65] as an adversarial loss with combination of $\ell 1$ loss in generator network. Table 3.1 shows the results based on different depth.

Regarding the highly imbalance datasets as shown in Figure 3.5, minority voxels with lesion label are not trained as well as majority voxels with non-lesion label. Therefore, we weighted only non-lesion classes to be in same average of lesion or tumor(s) classes. Table 3.2 and Table 3.4 describe our achieved results with and without weighting loss on BraTS 2017.

Configuration of recurrent-GAN

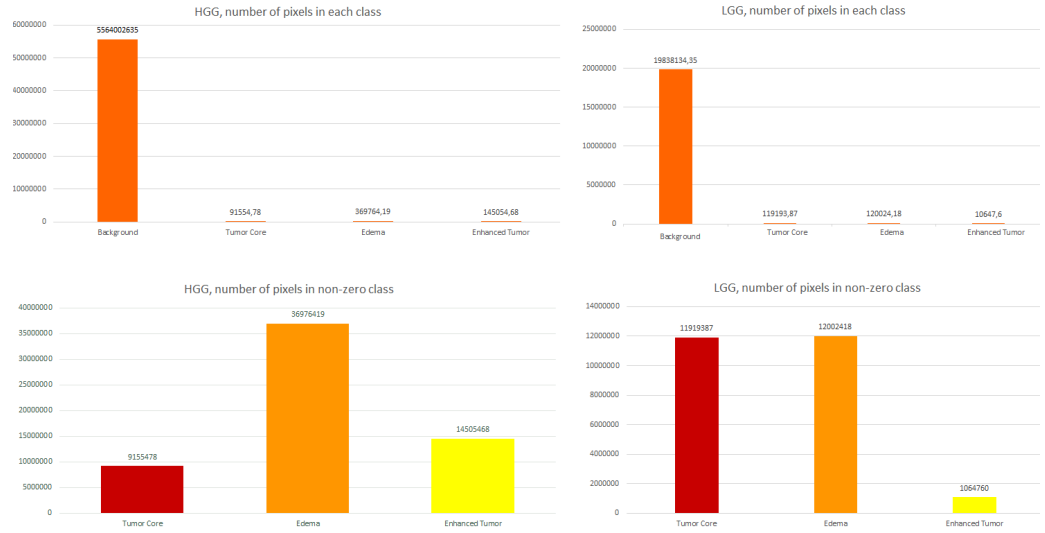


Figure 3.5: The number of pixels for each tumor classes represents how imbalanced is training data in detail of two subsets: high and low grade glioma brain tumor on BraTS 2018.

The recurrent-GAN architecture is implemented based on a Keras library [59] with back-end Tensorflow [62] and is publicly available on HPI-Deep Learning github [1]. training and experiments are conducted on a workstation equipped with a couple TITAN X GPUs. The learning rate is initially set to 0.0002. The RMSprop optimizer is used in both the generator and the discriminator, it dividing the learning rate by an exponentially decaying average of squared gradients. The model is trained for up to 120 epochs.

Network Architecture of recurrent-GAN

In this work, a generative network is a modified UNet architecture consist of ten fully convolutional layers, four max-pooling layers, dropout layer where a bidirectional LSTM unit is fed in the circumvent bottleneck. The bidirectional LSTM provides inter as intra-slice feature representation, which is very important in sequence medical image analysis. The advantage of bidirectional LSTM appears when we connected features from $n-1-i$ and i (as shown in Figure 3.2) from the same patient and passed them into the fully connected layer to classify patient

[1] https://github.com/HPI-DeepLearning/SV-GAN

diseases.

Our discriminator is 2D fully convolutional Markovian PatchGAN classifier [40] difference in bidirectional LSTM layers after the last fully convolutional layer. Unlike, the PathGAN discriminator introduced by Isola et al. [40] which classified each N $\times$ N patch for real or fake, we have achieved better results for the task of semantic segmentation in pixel level where we consider N=1. Moreover, since we have a sequential data, the bidirectional LSTM added after the last CNN layer in discriminator network. Same as encoder part of generator architecture, the discriminator consists of five fully convolutional layers and four max-pooling layers.

We used binary cross-entropy as an adversarial loss, categorical cross entropy [65] as an additional loss for the generative model for the task of semantic segmentation. The $\ell 1$ loss is for classification. Regarding the heavy class imbalance in both datasets, minority classes might not be trained as well as majority classes especially in the task of semantic segmentation, which we used selective weighted categorical cross entropy loss as segmentation loss. From Table 3.2 and Table 3.8, the second row provide results without weighted loss, while the first row in both Tables show the obtained results with selective weighted loss. In this work, the recurrent architecture selected for both discriminator and generator is a bidirectional LSTM proposed by Graves et al. [97].

3.3.3 Evaluation

We followed the evaluation criteria introduced by the BraTS 2018 [1], the ACDC2017 [1], the ISLES [2] challenge organizers. Moreover, we evaluate the proposed cost-sensitive losses by evaluation criteria regarding handling imbalanced issues.

Evaluation of voxel-GAN Network

The voxel-GAN is trained separately on BraTS 2017 and ISLES 2017 for semantic segmentation of lesion or tumor. The segmentation of the brain tumor

[1] http://www.med.upenn.edu/sbia/brats2018/evaluation.html
[2] https://www.smir.ch/ISLES/Start2018

Table 3.2: Comparison results of our achieved accuracy for semantic segmentation by voxel-GAN (trained model with weighted loss) with related work and top ranked team, in terms of Dice and Hausdorff distance on five fold cross validation after 80 epochs while the reported results in second and third rows are after 200 epochs. WT, ET, and CT are abbreviation of whole tumor, enhanced tumor, and core of tumor regions respectively.

	Dice			Hdff		
Methods	WT	ET	CT	WT	ET	CT
Voxel-GAN	0.84	0.63	0.79	6.41	7.1	10.38
cGAN [100]	0.81	0.61	0.64	7.30	9.22	12.04
Cycle-GAN [101]	0.90	0.78	0.81	2.50	4.5	5.4
Ensemble of 10 3D-Models [102]	0.91	0.82	0.86	3.9	4.5	6.8
3D UNet + TTA [103]	0.87	0.75	0.78	4.5	5.9	8.0

or lesion from medical images is highly interested in surgical planning and treatment monitoring. As mentioned by Menze et al. [2], the goal of segmentation is to delineate different tumor structures such as active tumor core, enhanced tumor, and whole tumor regions.

Figure 3.6 shows good trade-off between Dice and Sensitivities in training and validation time which it shows success for tackling of unbalancing data.

From obtained results on Table 3.2 and Table 3.3, the proposed voxel-GAN achieved better results in terms of Dice compared to 2D-cGAN. One likely explanation is that the voxel-GAN architecture is trained on 3D convolutional features, and the segmentor loss is weighted for imbalanced data.

Unlike previous works [101, 102, 103], we start training from scratch. From Table 3.2, two top ranked team used ensemble of pre-trained models. Ensemble networks provides good solution for imbalanced data by modifying the training data distribution with regards to the different misclassification costs. The qualitative results are shown in Figure 3.7.

We evaluated recurrent-GAN substituted by selective weighted loss for learning multi-task. We trained recurrent-GAN for simultaneous semantic segmentation and classification. The recurrent GAN is trained separately on ACDC 2017 and BraTS 2017.

Table 3.3: Comparison results of our achieved accuracy for semantic segmentation by voxel-GAN (trained model with weighted loss) with related work and top ranked team, in terms of sensitivity and specificity on five fold cross validation after 80 epochs while the reported results in second and third rows are after 200 epochs. WT, ET, and CT are abbreviation of whole tumor, enhanced tumor, and core of tumor regions respectively.

	Sen			Spec		
Methods	WT	ET	CT	WT	ET	CT
Voxel-GAN	0.86	0.74	0.78	0.99	0.99	0.99
cGAN [100]	0.75	0.61	0.55	0.99	0.99	0.99
Cycle-GAN [101]	0.89	0.89	0.81	0.99	0.99	0.99

Table 3.4: The achieved accuracy for semantic segmentation by 3D-GAN in terms of Dice and Hausdorff distance after 80 epochs. Here, the model trained based on 3D UNet as segmentor and 3D fully convolution as discriminator without weighting cost. The WT, ET, and TC are short of whole tumor, enhanced tumor, and tumorous core respectively.

Label	Dice-ET	Dice-WT	Dice-TC	Hdff-ET	Hdff-WT	Hdff-TC
Mean	0.438	0.633	0.481	54.2	12.9	33.70
StdDev	0.27	0.25	0.27	116.71	14.9	78.4
Median	0.48	0.73	0.57	8.76	8.0	11.70
25quantile	0.19	0.49	0.27	4.41	5.56	7.9
75quantile	0.65	0.82	0.70	20.82	14.08	19.1

Table 3.5: The achieved accuracy for semantic segmentation on ISLES dataset by voxel-GAN and conditional-GAN in terms of Dice, Hausdorff distance, average precision, and average recall on five fold cross validation after 200 epochs.

	Dice	Hausdorff	Precision	Recall
voxel-GAN	0.83	9.3	0.81	0.78
cGAN	0.75	14.6	0.74	0.73

(a)

(b)

Figure 3.6: The achieved accuracy obtained by voxel-GAN in terms of Dice and sensitivity at training and validation time on BraTS-2018.

Evaluation of recurrent-GAN network

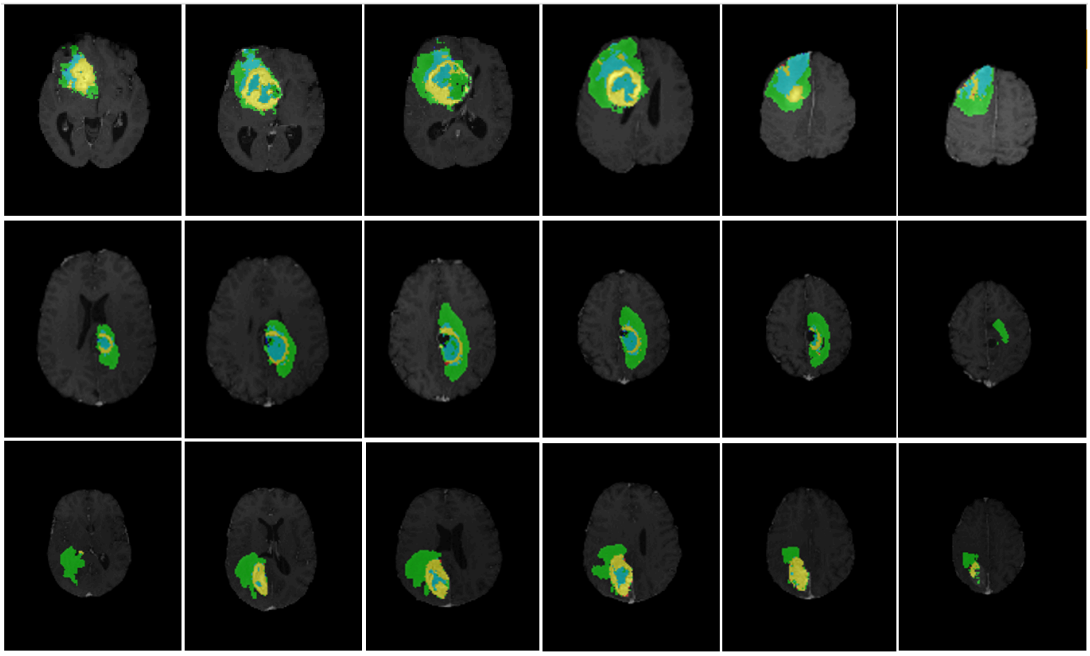


Figure 3.7: Predicted results from voxel-GAN model on axial views of Brats17-2013-37-1, Brats17-CBICA-AAC-1, and Brats17-CBICA-AAK-1 from the test set overlaid T1C modality. The green color codes the whole tumor (WT) region, while blue and yellow represent the enhanced tumor (ET) and the tumorous core (TC) respectively.

Heterogeneous Brain Tumor Segmentation and Diseases Classification The goal is simultaneous semantic segmentation of different tumor structures such as active tumorous core, enhanced tumorous, and whole tumorous and identifying brain diseases (see Table 3.6). More than 80 teams attended to the BraTS 2017 challenge which Table 3.2 shows comparison results between most related approaches and top three groups obtained by the organizer [104].

From Table 3.2, our the proposed method achieved 8% better results for whole tumor segmentation in terms of Dice, compared to the conditional GAN [100] where the UNet is a generator, and fully convolutional network is discriminator. One likely explanation is that our method trained on the proposed weighted loss where we attenuate the effect of imbalanced data. Another reason is our architecture where semantic segmentation training is from 2D sequence medical images into a 2D sequence of semantic segmentation. As mentioned before, both generator and discriminator models are designed to be aware of contextual features

Table 3.6: The achieved accuracy for classification of brain diseases by proposed multi tasks conditional GAN and comparison with related approaches.

Method	Dice	Precision	Recall
Multi tasks RNN-GAN	0.96 ± 0.03	0.9 ± 0.16	0.9 ± 0.21
Random Forest [105]	0.96	0.94	0.94
SVM [106]	0.87	0.94	0.94

from previous slices and able to pass aggregated features to next slices. In other words, the network learned interslice features as well as intra-slice features, which is a critical point in medical image analysis. Figure 3.8, shows predicted semantic segmentation region in three different patients from the local test set.

Based on Table 3.2, we see the comparison and performance of our method in detail with other related approaches as well as winners in terms of Dice, Hausdorff distance, and sensitivity. Based on Table 3.2, Deep Medic [107] achieved best results for whole tumor segmentation with the two cascade architecture of 3D U-Net and 3D-FCN and applying variation intensity normalization in each network. The UCL-TIG [108] achieved the second rank of a challenge with a triple cascade of 2D U-Net architecture where each network is trained on three different regions of the tumor.

Similar to the winners of BraTS 2017 [107, 108, 110], we do tumorous segmentation by attenuating imbalanced data effect, where they trained models in cascade architectures but applied weighted loss function. Compared to [107, 108, 110], our proposed network has the advantage of carrying out multiple clinical tasks in a single architecture where we have achieved 98.61% accuracy for diagnosis between high and low-grade glioma tumor, the performance of classification considering other metrics reported in Table 3.6. It is important to mention that our method takes only 58 seconds to segment one MR brain image consisting of 155 slices at testing time.

Cardiac Image Semantic Segmentation and Diseases Prediction

Cardiac function is important for the diagnosis and treatment of heart failure with infarction such as dilated cardiomyopathy, hypertrophic cardiomyopathy,

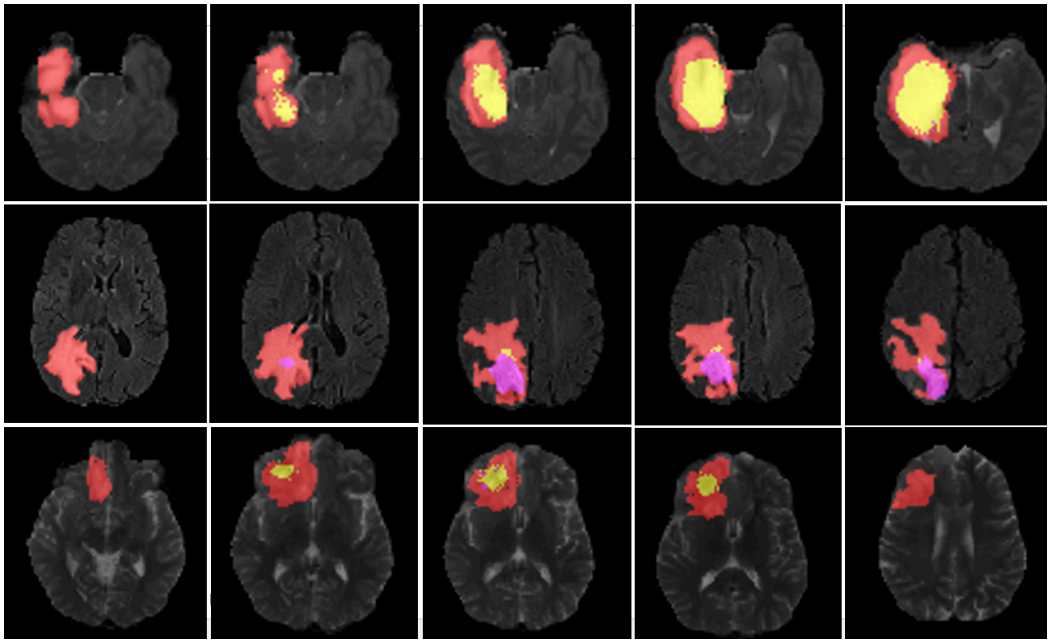


Figure 3.8: Segmentation results obtained by our proposed method which in the first row, the red, pink, and yellow colors respectively show the whole tumorous, enhanced region and the active core of tumor overlaid on the Flair modality.

and abnormal right ventricle. Cardiac magnetic resonance imaging (CMRI) provides a non-invasive diagnosis tool to study cardiac anatomy. Automating segmentation of cardiac images plays an important rule and desire application in a routine clinical task. In this regard, the ACDC [1] organized challenges for fully or semi-automatic approaches based on deep learning in two sections of cardiac diseases prediction and cardiac semantic segmentation. Our prediction results and comparison with recent related approaches are reported in [1] and described in Table 3.8.

Our predicted results (Figure 3.9 and Table 3.8), from test time show good relation to the ground truth for the left ventricle at the end of diastolic time. The good results for left ventricle expected because of the nature of LV shape. The average value of the Dice index is around 0.93 for myocardium vessel, which is slightly better than the result of the ACDC-2016 challenge winner [1]. The primary source of error here is the inability of the method to completely segment the right ventricle since it has a complex crescent shape across slices and phases (ED, ES).

Table 3.7: The achieved accuracy for semantic segmentation by proposed method in terms of Dice, Hausdorff distance (Hdff), and Sensitivity (Sen) on unseen data and comparison with related and top rank approaches. The WT, ET, and TC columns respectively are abbreviation of whole tumorous region, enhanced tumorous, and tumorous active core.

Segmentation	Dice			Hdff			Sen		
Architecture	WT	ET	TC	WT	ET	TC	WT	ET	TC
Weighted RNN GAN	0.88	0.76	0.77	6.11	7.17	11.38	0.87	0.88	0.85
cGAN	0.80	0.61	0.61	7.30	9.22	12.04	0.75	0.61	0.55
SegAN [41]	0.85	0.66	0.70	00	00	00	0.80	0.62	0.65
3D UNet [109]	0.88	0.72	0.76	13.6	13.8	22.3	00	00	00
3D UNet [107]	0.90	0.74	0.79	4.23	4.50	6.56	0.89	0.78	0.86
Cascade 2D UNet [108]	0.90	0.78	0.83	3.89	3.28	6.48	0.91	0.77	0.82
2D3D ResUNet [110]	0.89	0.74	0.80	6.97	4.55	9.48	0.89	0.79	0.78

As depicted in Figure 3.9 and Table 3.8, a right ventricle is the most challenging organ to segment where the most failure happened in the systolic phase. Based on Figure 3.9, the achieved accuracy in the test time on ACDC benchmark, we observed the average results in diastolic phase (sixth, seventh, and eighth columns) are better than the average results on systolic phase (third, fourth, and fifth columns).

Regarding the results from Table 3.10, our achieved accuracy in test time for classification of heart disease is 94%, which is promising results; especially for clinical application. Obtained results in Table 3.8 and Table 3.10 show the ability of the proposed approach for routine clinical application in two tasks of cardiac segmentation and heart failure prediction. Moreover, our method takes only 14 seconds to segment one cardiac image consisting of 20 slices at testing time, which is a crucial aspect in clinical application.

Table 3.8: Comparison and achieved accuracy in term of Dice metric and Hausdorff distance in detail of the end of systolic (ES) and end of diastolic (ED) phase from ACDC benchmark with related approaches and top-ranked methods reported in [1]. The LV, RV, and MYO columns respectively are the abbreviation of left ventricle region, right ventricle region, and myocardium vessel.

Segmentation		Dice-ES			Dice-ED	
Architecture	LV	RV	MYO	LV	RV	MYO
Weighted-RNN-GAN	0.935	0.921	0.926	0.967	0.947	0.930
cGAN by ours	0.918	0.874	0.870	0.930	0.902	0.899
Dilated CNN [74]	0.918	0.872	0.894	0.968	0.928	0.875
UNet [76]	0.905	0.882	0.896	0.957	0.941	0.884
2D M-Net [111]	0.921	0.885	0.895	0.959	0.929	0.884
2D UNEt [112]	0.911	0.883	0.901	0.963	0.932	0.892
2D+ 3D UNet [77]	0.931	0.899	0.919	0.968	0.946	0.902
SVF-Net [75]	0.900	0.845	0.869	0.957	0.916	0.869
3D UNet [113]	0.775	0.770	nr	0.864	0.789	nr

Table 3.9: Comparison and achieved accuracy in term of Dice metric and Hausdorff distance in detail of the end of systolic (ES) and end of diastolic (ED) phase from ACDC benchmark with related approaches and top-ranked methods reported in [1]. The LV, RV, and MYO columns respectively are an abbreviation of left ventricle region, right ventricle region, and myocardium vessel.

Segmentation		Hdff-ES			Hdff-ED	
Architecture	LV	RV	MYO	LV	RV	MYO
Weighted-RNN-GAN	8.02	12.08	8.65	6.81	8.97	8.04
cGAN by ours	9.44	13.04	9.5	8.60	8.95	8.08
Dilated CNN [74]	9.6	13.4	10.7	7.5	11.9	11.1
UNet [76]	8.7	14.1	9.3	6.6	10.3	8.7
2D M-Net [111]	7.1	11.8	8.9	7.7	12.9	9.9
2D UNEt [112]	9.2	14.5	10.6	6.5	12.7	8.7
2D+ 3D UNet [77]	6.9	12.2	8.7	7.4	10.1	8.7
SVF-Net [75]	10.9	15.9	13.03	7.5	14.1	11.5
3D UNet [113]	53.1	31.1	nr	47.9	30.3	-

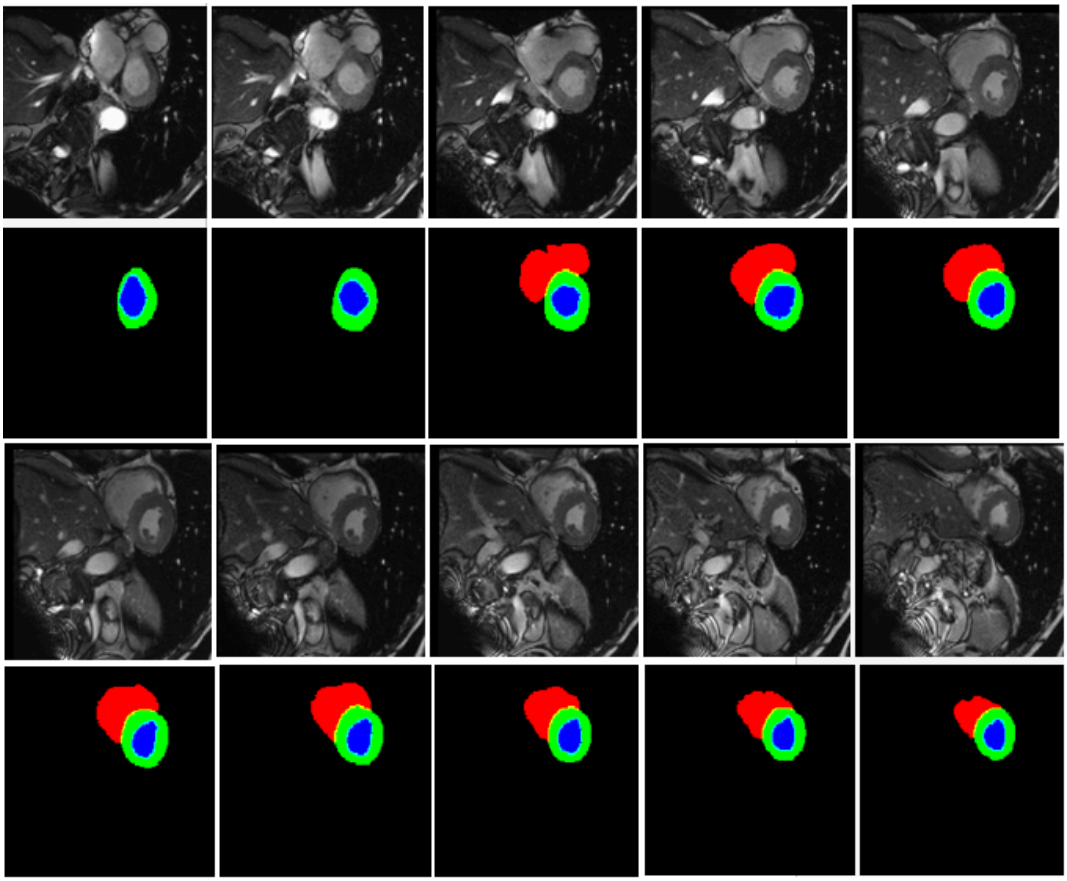


Figure 3.9: The cardiac MR images in systolic phase from t=0 till t=9 in the top row and second row represent the segmentation results obtained by our proposed method from ACDC 2017 benchmark on Patient084 where the red, green, and blue contour present respectively right ventricle, myocardium, and left ventricle region.

Table 3.10: The achieved accuracy for classification of cardiac diseases by proposed multi tasks conditional GAN and comparison with related approaches and top-ranked obtained by ACDC reported in [1].

Method	Accuracy	Dice	Precision	Recall
Multi tasks RNN-GAN	0.94 ± 0.16	0.94 ± 0.16	0.94 ± 0.16	0.94 ± 0.16
Random Forest [105]	0.96	0.94	0.94	0.94
SVM [106]	0.87	0.94	0.94	0.94
Random Forest [77]	0.92	0.94	0.94	0.94
Random Forest [74]	0.85	-	-	-

Chapter 4

Approaches to Handle Imbalanced Data through Class Expert Ensembles

Ensemble methods use multiple learning algorithms to obtain better predictive performance than an individual model. The key idea of this chapter is to provide better representation on an imbalanced dataset using various deep learning architectures or different classifiers as an ensemble model.

In this chapter, we introduce three different ensemble architectures aim to mitigate the imbalanced problem in the task of semantic segmentation. Conventional GAN comprises two models: a generative model and a discriminative model. The generative probabilistic model builds the model based on prior domain knowledge about the appearance and spatial distribution of the different tissue types while the discriminative model directly learns the relationship between the local features of images and labels. The training procedure for the generative and the discriminative is similar to a two-player mini-max game, where a generative a discriminative are trained in an alternating fashion to minimize and maximize an objective function respectively. Hence, an inevitable discriminative loss can reduce the error of prediction of the generative model regarding some aspects of quality. Therefore, we study the impact of generative ensemble discriminative networks on imbalanced image semantic segmentation, considering various losses and different architectures.

In our work, we have investigated three different ensemble models including conditional generative adversarial networks based on stacking ensemble, bagging ensemble GANs, and cascade GANs which differs from the type of ensembles, image representation (2D and 3D), architectural design, and losses.

We demonstrate the resulting segmentation algorithms on popular medical imaging benchmarks for segmentation of abnormal tissues as well as anatomical organs. We close the chapter with ongoing directions and insights about imbalanced learning by ensemble models.

4.1 Technical Background

Ensemble Learning

Ensemble learning generates multiple classifiers and combines them to a single consensus model to solve a learning problem. The key idea is that a consensus model performs better than an individual model, or it can reduce the likelihood of selecting a model with inferior performance as shown by Polikar [114]. Each classifier of the ensemble model might be predicting class labels, posterior probabilities, real-valued number, clustering, or any other quantity. Therefore their decisions can be combined by many methods such as voting, averaging, and probabilistic approach.

The core idea on ensemble learning is collective wisdom (or the wisdom of crowds) that many heads are smarter than the few or one. As discussed by James Surwiecki [115], not all crowds are wise, but to become wise, the crowd should comply with the diversity of opinion and other criteria. We expected each model in an ensemble framework might have an individual element of *diversity* to retaining good performance. We can roughly divide the existing ensemble methods into two categories by considering the element of diversity [116]: those that encourage diversity *implicitly*, and those that encourage *explicitly*.

Implicit is the most popular ensemble models where different random subsets of the training data are assigned to each learner. Diversity is encouraged *implicitly* by random sampling of the data space: at no point is a measurement taken to ensure diversity will emerge. The random differences between the datasets might be in the selection of examples (the Bagging algorithm [117]), the selection of

features (Random Subspaces [118]), or combinations of the two (the Random Forests [119] algorithm). The propose cascade-GAN (see Figure 4.1-a) is an implicit ensemble model which different batches train on each stage of GAN.

Alternatively, *explicit* model encourage diversity to construct each ensemble member with some measurement ensuring it is substantially different from the other members. Boosting algorithms [120] achieve this by modifying the distribution of training samples for each learner. It is encouraged to make more accurate predictions where previous predictors have made errors.We consider, ensemble GAN (see Figure 4.1-c) including one generator and k multiple discriminators which we study the effect of various losses and different architectures for handling imbalanced class problem in semantic segmentation.

Imbalanced Learning using Ensemble Model

In the context of imbalanced data, the main idea of combining multiple classifiers in redundant ensembles is to improve their generalization ability. Each classifier can cover only some aspects of the quality of the application. Hence, the patterns that are misclassified by different classifiers are not the same, then combining classifiers to a single consensus model performs better than one classifier.

Form statistical perspective, an ensemble model composed by diverse classifiers is studies in terms of bias-variance decomposition [121] and bias-variance-covariance-variance noise decomposition [122]. The bias is the difference between the average prediction of the model and the real value, or ability of the model to generalize correctly on the testing set. The variance is the variability of model prediction for a given data point or sensitivity of the model to small fluctuations in the training set.

Hence variance is related to overfitting, the performance improvement in the ensemble model is due to the reduction in variance because the natural effect of ensemble averaging is to reduce the variance of a set of classifiers. On the other side, ambiguity decomposition shows the combination of several classifiers is better on average of several patterns than a single classifier. Some [123] has been focused on multiple classifiers as a regression problem where the output is real-valued, and the mean squared error is used as the loss function. However, in the context of classification, those terms are still [124], since different authors provide

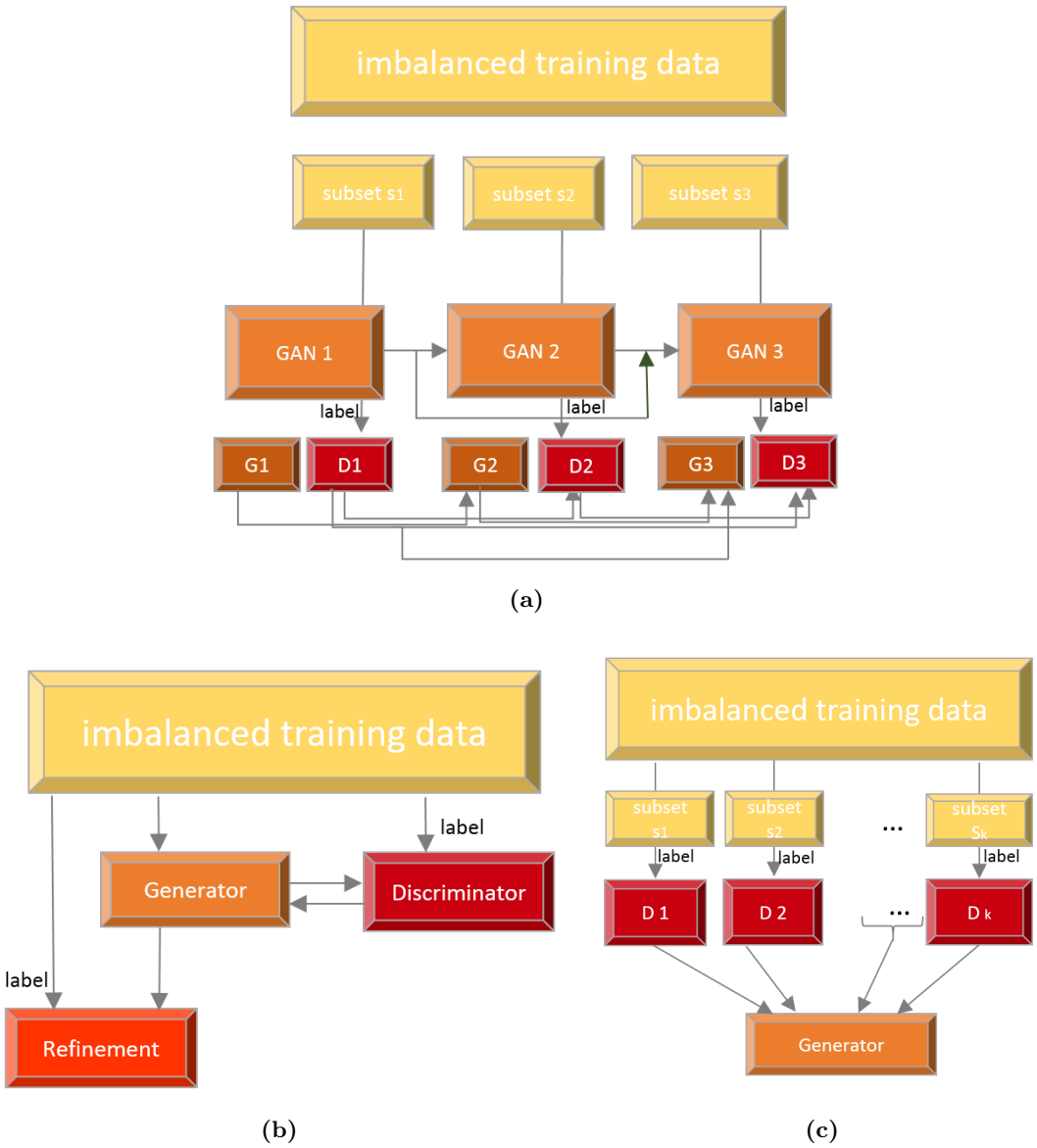


Figure 4.1: GAN-based ensemble architectures for handling class imbalanced problem. (a) cascade-GAN, (b) conditional refinement GAN, (c) generative ensemble discriminative.

different assumptions [125] and there is no an agreement on their definition for generalized loss functions [126].

Ensemble models are popular and most successful model in medical imaging which recently these methods achieved the best performance number of challenges; such as 2018 BraTS, Myronenko. [127] proposed a deep ensemble archi-

tecture consisting ten models, in 2016 Camelyon challenge, Wang et al. [128] applied ensemble of two GoogLeNet architectures, one trained with and one without hard-negative mining to tackle the challenge, Schaumberg et al. [129] achieved the best performance for prostate cancer detection using an ensemble of ResNet architecture. Gunhan Ertosun and Rubin. [130] applied ensemble of CNNs for grading glioma, Ciresan et al.[131] applied several different architectures for large organ segmentation. In this chapter, we introduce three different ensemble models to deal with the imbalanced class problem in semantic segmentation, as shown in Figure 4.1.

4.2 Ensemble Model

In this section, we describe our three different ensemble networks to mitigate the imbalanced problem in the task of medical image semantic segmentation. The proposed ensemble methods are based on generative adversarial networks differs on architectural design, losses, image representation, and type of ensembles.

4.2.1 Conditional Generative Refinement Network

We propose a conditional generative refinement network with three components: a generative, a discriminative, and refinement networks to mitigate imbalanced data problem through ensemble learning. The generative network learns to the segment at the pixel level by getting feedback from the discriminative network according to the true positive and true negative maps. On the other hand, the refinement network learns to predict the false positive and the false negative masks produced by the generative network that has significant value, especially in medical application. The final semantic segmentation masks are then composed by the output of the three networks. Here we build an ensemble framework by training each new model instance to learn to address misclassified samples by previous model same as boosting strategy. Boosting tends to improve upon its base models.

In conventional generative adversarial networks, generative model G tries to learn a mapping from random noise vector z to output image y; $G : z \rightarrow y$. Meanwhile, a discriminative model D estimates the probability of a sample coming

from the training data x_{real} rather than the generator x_{fake}. The GAN objective function is a two-player mini-max game like Eq.(4.1).

$$\min_G \max_D V(D,G) = E_y[logD(y)] + E_z[log(1 - D(G(z)))] \tag{4.1}$$

In a conditional GAN, a generative model learns the mapping from the observed image x and a random vector z to the output image y; $G : x, z \rightarrow y$. Discriminative model on the other hand attempts to discriminate between generator output and ground truth of the training set. Unlike previous conditional GANs [11, 40, 41, 42, 43]; in our proposed method, a generative model learns mapping from a given sequence of 2D multimodal MR images x_i to a sequence semantic segmentation y_{seg}; $G : \{x_i, z\} \rightarrow \{y_{seg}\}$ (where i refers to 2D slice index between 1 and 155 from a total 155 slices acquired from each patient). We utilize bidirectional LSTM to pass the temporal consistency between 2D slices. Our network can learn representations from previous and future slices, which results in context-aware and eliminate ambiguity. The training procedure for the segmentation task is similar to two-player mini-max game, as shown in Eq.(4.2). While the generator segmented pixels label, the discriminator takes the ground truth, and the generator's output to classify whether the output is real or fake.

$$\begin{aligned}\mathcal{L}_{adv} \leftarrow \min_G \max_D V(D,G) = E_{x,y_{seg}}[logD(x, y_{seg})] + \\ E_{x,z}[log(1 - D(x, G(x,z)))]\end{aligned} \tag{4.2}$$

Here, the generative loss Eq.(4.3) is mixed with $\ell 1$ term to minimize the absolute difference between the predicted value and the existing largest value. Previous studies [40, 41] on cGANs have shown the success of mixing the cGANs objective with $\ell 1$ distance. The $\ell 1$ objective function takes into account CNNs feature differences between the predicted segmentation and the ground truth segmentation and resulting in fewer noises and smoother boundaries.

$$\mathcal{L}_{L1}(G) = E_{x,z} \parallel y_{seg} - G(x,z) \parallel \tag{4.3}$$

The adversarial loss for semantic segmentation task calculate by Eq.(4.4)

$$\mathcal{L}_{seg}(D,G) = \mathcal{L}_{adv}(D,G) + \mathcal{L}_{L1}(G) \tag{4.4}$$

Figure 4.2: Visual results from our model where the cGAN over segment through learning true positives and true negatives and the refinement learns false positives and false negatives mask.

In order to tackle with misclassification cost, the predicted output by the generator and discriminator are passed to refinement network. The refinement network is trained to learn the false prediction of cGAN in details of false negatives (same as Eq. 4.5) and false positives (same as Eq. 4.6). The false negative error represents the number of pixels that were incorrectly labeled as background or wrong class (Figure 4.2-third column). Similarly, the false positive indicates the number of pixels that were incorrectly labeled as part of the region of interest (Figure 4.2-last column).

$$\mathcal{L}_{fn} = clip((y - \mathcal{L}_{seg}), 0, 1) \tag{4.5}$$

$$\mathcal{L}_{fp} = clip((\mathcal{L}_{seg} - y), 0, 1) \tag{4.6}$$

where in both equations (4.5 and 4.6) y, $\mathcal{L}_{seg}$ respectively refers to the ground truth labels and predicted labels by adversarial loss.

Our final objective function $\mathcal{L}_{CR-GAN}$ for semantic segmentation relies on adding false negatives and subtracting false positives from the adversarial network predicted output.

$$\mathcal{L}_{CR-GAN} = \mathcal{L}_{seg} - \mathcal{L}_{fp} + \mathcal{L}_{fn} \tag{4.7}$$

As shown in Figure (4.3), our proposed method consists of a generator network, and a discriminator network, in the left side, followed by a refinement network in the right side of the figure. We investigate two different architectures of conditional GAN and recurrent conditional GAN for adversarial training of G and D.

Conditional Generative Adversarial Network

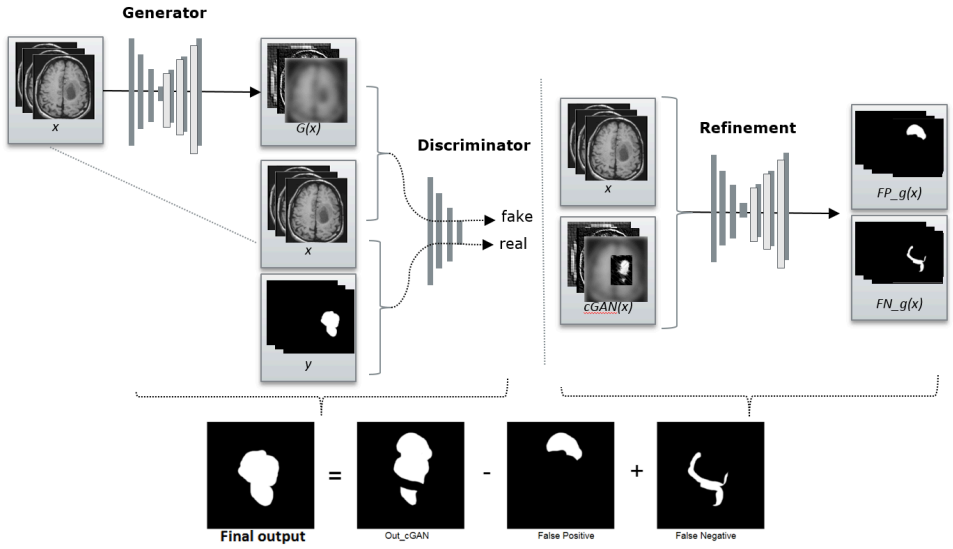


Figure 4.3: The proposed method for medical image semantic segmentation consists of a generator, a discriminator, and refinement networks. The generator tries to segment image in pixel level, while discriminator classifies the synthesized output is real or fake. The final semantic segmentation masks are computed by eliminating the false positive and adding the false negative predicted masks by the refinement network.

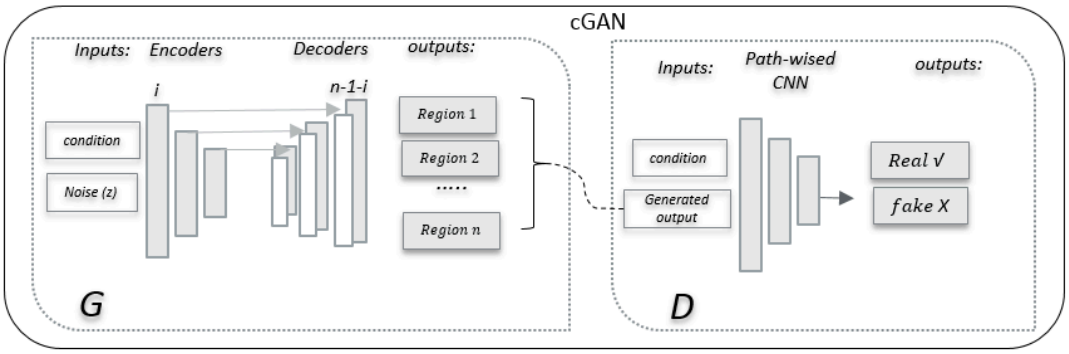


Figure 4.4: Conditional generative adversarial networks, consists of a generative model and a discriminative model where can be constructed by feeding the data, we wish to condition on to both the generative and discriminative.

As depicted in Figure 4.4, the conditional GAN composed by a generator and a discriminator where the generator is a fully convolutional encoder-decoder network that generates a label for each pixel. Similar to UNet [66], we added the skip connections between each layer i and layer $n-i$, where n is the total number of layers. Each skip connection concatenates all channels at layer i with those at

layer $n-i$. We use the convolutional layer with kernel size 5×5 and stride 2 in encoder part for down-sampling, and in the decoder, section performs up-sampling by image re-size layer with a factor of 2 and convolutional layer with kernel size 3×3 stride 1. In our architecture, in the last layer, the high-resolution features from multi-modal, multi-site images are concatenated with up-sampled versions of global low-resolution features, which helps the network learn both local and global representation of features.

The discriminator is a fully convolutional network and has the same architecture as decoder part of the generator network. The hierarchical features from convolutional layers passed to softmax loss for classifying whether a segmented pixel's label belongs to the right class.

Recurrent Generative Adversarial Networks

The recurrent cGAN framework consists of ten generators and ten discriminators which the generator are connected through auto-encoder's bottleneck that substituted with bidirectional LSTM units [97]. The discriminators are connected by bidirectional LSTM units [97] in the last layer. Similar to few-shot learning, we train the network only with few slices. The recurrent conditional GAN has the advantage of getting the consistency information between previous and next slice. Using bidirectional LSTM units inside of G and D makes networks context-aware, which is a crucial point in sequence data analysis.

More important, annotated data obtained by medical expert will not always be possible and are rare. Therefore training network in this way is more applicable and welcome for routine clinical task.

Refinement Network

In order to address misclassified samples, we design the refinement network on top of an adversarial network to deal with unbalanced data issues and improve false positive rate. As shown in Figure 4.5, the refinement network is the fully convolutional networks, more specifically, similar to UNet style with bidirectional LSTM in circumventing of a bottleneck. The refinement network takes a 2D sequence outputs from cGAN (or recurrent cGAN), with a 2D sequence of medical

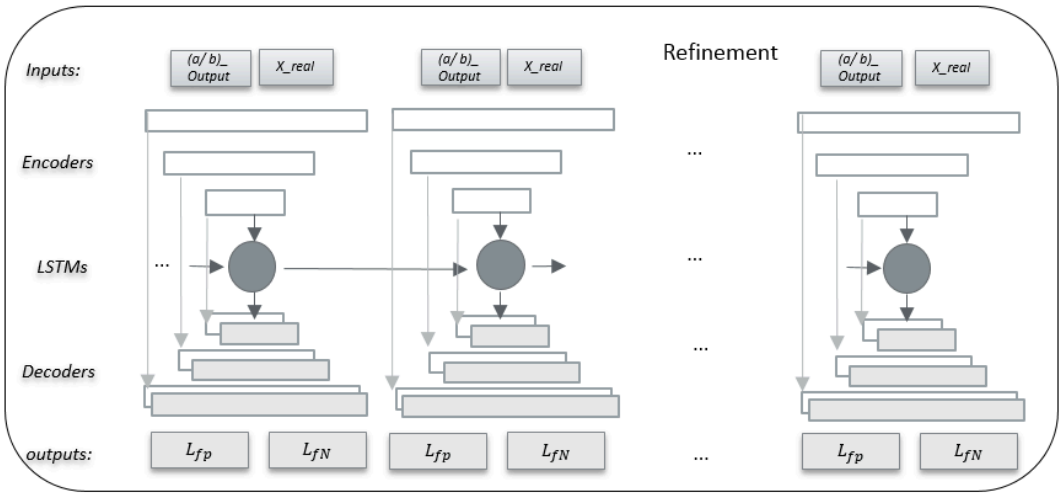


Figure 4.5: the proposed refinement network consist of fully convolutional encode-decoder.

images, and the outputs are a 2D sequence masks of false positives and false negatives.

The final semantic segmentation calculated by adding false negatives and subtracting false positives mask to the output of cGAN network. We gather all the described networks, in a single framework and train a cGAN network with a refinement network end-to-end. Similar to recurrent-GAN setting, the refinement is trained with few-slices.

Algorithm 2: The conditional Refinement GAN for medical image semantic segmentation

Input : Sequence of 2D medical images from plane x, y, z

Output: Sequence of 2D binary masks from plane x, y, z

First stage: Automatic Segmentation by cGAN [11]:

for $i = 1, D_{steps}$ **do**

  Sample patient-wise mini-batch, noise samples z, from noise prior $p_g(z)$

  Sample patient-wise mini-batch, from data generating distribution $p_{data}(x)$

  Update the discriminator following:

  $Maximizing(E_x \sim p_d[logD(x; \theta_d \mid c)] + E_z \sim p_z(z)[log(1 - D(G(z; \theta_g \mid c); \theta_d))])$

end

for $i = 1, G_{steps}$ **do**

  Sample patient-wise mini-batch, noise samples z, from noise prior $p_g(z)$

  Update the generator (segmentor)following:

  $Minimizing(E_z \sim p_z(z)[log(1 - D(G(z; \theta_g \mid c); \theta_d))])$

end

Second stage: Automatic Refinement Segmentation:

for $i = 1, Refinement_{steps}$ **do**

  Sample patient-wise mini-batch, from first step of conditional generating distribution $p_{data}(x)$

  Optimize the error of conditional generating distribution and true data following:

  $\mathcal{L}_{fn} = clip((y - \mathcal{L}_{GAN}), 0, 1)$

  $\mathcal{L}_{fp} = clip((\mathcal{L}_{GAN} - y), 0, 1)$

end

final objective function calculated as follow:

$\mathcal{L}_{Seg} \leftarrow (\mathcal{L}_{cGAN}) - \mathcal{L}_{fp} + \mathcal{L}_{fn}$

4.2.2 Cascade of Generative Adversarial Networks

We proposed cascade of the conditional generative adversarial network consists of three individual GAN frameworks. These three frameworks are trained separately on different stages where each stage designed to share convolutional features and weights, where the later stages use the shared convolution features from the previous stage and transfer the learned convolutional features and weights to the

next stage. Each stage involves a cGAN with an adversarial loss and individual parameters.

The GAN framework on each step has two networks of a generator and a discriminator. We train the discriminator model D and the generator model G simultaneously in an adversarial way. The generator model G maps the pixel's label of a specific image and D tries to distinguish the predicted image comes from the reference distribution or the generative network.

$$\mathcal{L}_{GAN}(G, D) = E_y \; p_{data(y)}[logD(y)] + E_x \; p_{data(x)}, z \; p_z(z)[log1 - D(x, z)] \quad (4.8)$$

Generator Architecture of cascade-GAN

As shown by Figure 4.6, the generator is a fully convolutional encoder-decoder structure like a UNet with skip connections between corresponding layers in the encoder and the decoder. We use the convolutional layer with kernel size 4×4 and stride 2 for down-sampling, and perform upsampling by image re-size layer with a factor of 2 and the convolutional layer with kernel size 3×3 stride 1.

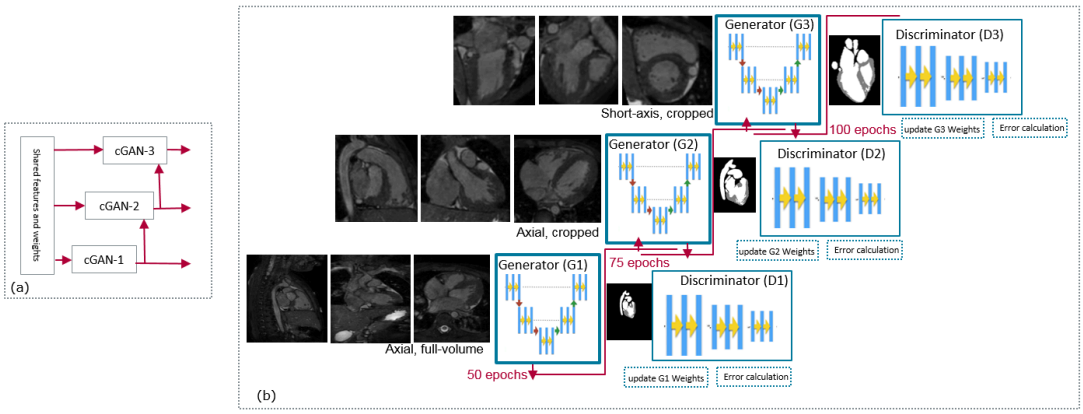


Figure 4.6: The cascade-GANs with three stages (a). The proposed architecture is context-aware where the later stages use the shared convolution features plus the probability map obtained from previous stages and transfer the learned convolutional features to the next (b).

Discriminator Architecture of cascade-GAN

The discriminator is a fully convolutional neural network. Hierarchical features are extracted from multiple layers of convolution and used to compute the $\mathcal{L}1$ loss function. We confirmed the solution suggested by Isola et al. [132] for using $\mathcal{L}1$ distance rather than $\mathcal{L}2$ as $\mathcal{L}1$ encourages less blurring. $\mathcal{L}1$ loss can capture long- and short-range spatial relations between pixels by using the hierarchical features.

Both the generator G and the discriminator D are trained through back propagation from the proposed $\mathcal{L}1$ loss. The training of the generator and the discriminator is like playing a mini-max game Eq.(4.8): While the goal G is to maximize the discriminator loss, D tries to minimize it. This training process makes both networks increasingly powerful.

Training Procedure by cascade-GANs

As shown in Figure 4.6-a, in the cascade-GANs, later stages use the shared convolution features plus the probability map obtained from previous stages and transfer the learned convolutional features to the next. We train iteratively several GANs that take as input MRI patches and estimate corresponding segmented binary patches. These patches are concatenated as a second channel in the MRI patches, and this new data is used as input during the training of the next GAN. An illustration of this scheme is shown in Figure 4.6-b.

Algorithm 3: The context-aware conditional generative adversarial networks for medical image semantic segmentation

Input : minibatch samples b from 2D medical images

Output: 2D segmented masks

For each stage: Automatic Segmentation by cGAN [11]:

for $i = 1, D_{steps}$ **do**

  Sample mini-batch b, noise samples z, from noise prior $p_g(z)$

  Sample from data generating distribution $p_{data}(b_x)$

  Update of finetune the discriminator following:

  $Maximizing(E_x \sim p_d[logD(b_x; \theta_d \mid c)] + E_z \sim p_z(z)[log(1 - D(G(z; \theta_g \mid c); \theta_d))])$

end

for $i = 1, G1_{steps}$ **do**

  Sample patient-wise mini-batch, noise samples z, from noise prior $p_g(z)$

  Update and fine tune the generator following:

  $Minimizing(E_z \sim p_z(z)[log(1 - D(G(z; \theta_g \mid c); \theta_d))])$

end

4.2.3 Ensemble-GANs

We present a new generative ensemble adversarial discriminative network that can effectively tackle the imbalanced problem in semantic segmentation. Our idea is to use a mixture of many discriminator losses rather than a single one in the conventional GAN, to cover and minimize the prediction error of semantic segmentation by generator from different aspects of quality.

Figure 4.7 illustrates the architecture of our proposed generative ensemble network, where all components are parameterized by neural networks. In our workflow, the generator G is forced and learned to minimize the prediction error of semantic segmentation through the ensemble of discriminators. This ultimately encourages G to produce conditional samples with minimum error, since G needs to fool the different possible discriminators. Variations in the ensemble are achieved by the feedback of each D with a certain probability at the end of every batch. This means that G will only consider the loss of the remaining discriminators in the ensemble while updating its parameters at each iteration.

Algorithm 4: The generative ensemble adversarial discriminative networks for medical image semantic segmentation

Input : Training samples $S_N := X_1, X_2, ..., X_N$

Output: Generated Samples by generative model

for $numberofiterations$ **do**
  Sample mini-batch from training sample $G_N := X_1, X_2, ..., X_N, x_i \sim p_{g_{data}}(x)$
  Sample mini-batch from Gaussian noise $z_i := z_1, z_2, ..., z_N, z_i \sim p_g(z)$
  for $k = 1, D_{k_{steps}}$ **do**
    Sample mini-batch b, noise samples z, from noise prior $p_g(z)$
    Sample from data generating distribution $p_{data}(b_x)$
    Update of fine-tune k discriminators following:
    $Maximizing(E_x \sim p_d[logD_k(b_x; \theta_d \mid c)] + E_z \sim p_z(z)[log(1 - D_k(G(z; \theta_g \mid c); \theta_d))])$
  end
  for $i = 1, G_{steps}$ **do**
    Sample patient-wise mini-batch, noise samples z, from noise prior $p_g(z)$
    Update and fine-tune the generator following:
    $Minimizing(E_z \sim p_z(z)[log(1 - D(G(z; \theta_g \mid c); \theta_d))])$
  end
end

We explore a single generator with k different discriminators: (1) a more discriminating with different losses are able to cover more aspects of qualify for generator's output by approximating $max_D V(D, G)$; (2) a more discriminating with different representation of data are capable to better catch generator distributions.

$$\min_G \max_{D_k} V(D_k, G) = E_{x \sim p(x)}[logD_k(x)] + E_{z \sim p(z)}[log(1 - D_k(G(z)))] \quad (4.9)$$

4.2.3.1 Generative ensemble adversarial losses

We propose generative ensemble adversarial losses including a single generator which tries to minimize segmentation error regarding ensemble of k different

losses. The generator takes random vector z, and 3D medical images x as conditional input meanwhile discriminators tries to minimize error of predicting segmentation masks by generator through multiple losses. Here, for fixed G, average function F getting feedback from k different discriminator's losses to generator by objective of $min_G max F(V(D_1, G), V(D_2, G), ..., V(D_k, G)))$.

$$\min_G \max_{D_i} V(D_k, G) = E_{x \sim p(x)}[log D_k(x, y)] + E_{z \sim p(x,z)}[log(1 - D_k(x, G(x, z)))] \tag{4.10}$$

where $D_k(x)$ and $G(z)$ are the outputs of the kth discriminator and the generator, respectively. The idea of using the proposed averaging scheme is to privilege worse discriminators and thus providing more useful gradients to the generator during training.

We adopt the generator with 3D modified hour-glass network described in voxel-GAN [93] consisting of two, 3D fully convolutional encoder-decoder networks that predict a label for each voxel. The first encoder takes $128 \times 128 \times 32$ of multi-modal MRI or CT images at the same time as different channel input. Last decoder outputs 3D images with same input size, $128 \times 128 \times 32$. Similar to UNet [66], we added the skip connections between each layer i and layer $n - i$, where n is the total number of layers in each encoder and decoder part. Each skip connection simply concatenates all channels at layer i with those at layer $n - i$. Moreover, we concatenate the bottleneck features and last convolutional decoder to capture better feature representation.

The discriminators are 3D fully convolutional encoder network which classifies whether a predicted voxel label belongs to the right class. More specifically, each discriminator is trained to minimize the error of predicted mask by generator regarding different losses. Experiments were performed for generator in conditional setting and by considering $k = 1, 2, 3$ as number of discriminators. However, results showed that the simple average of discriminators losses provided the better trade-off between precision and recall and improve segmentation results. More detail about different losses and and evaluation is provided in Section 4.3.

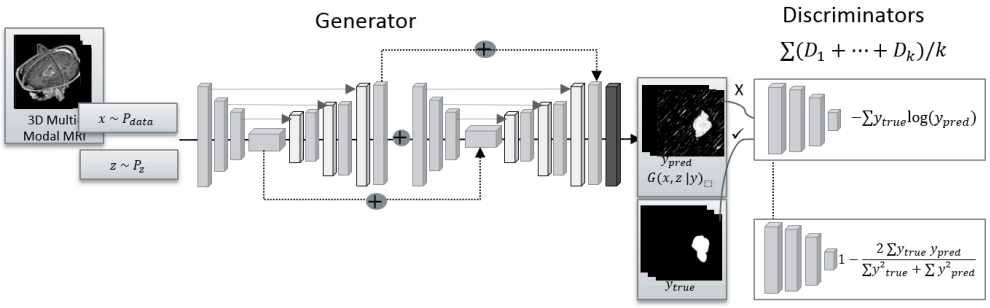


Figure 4.7: The proposed generative adversarial ensemble discriminators

4.3 Experimental Results

To evaluate the performance of our different implemented ensemble framework on handling imbalanced image semantic segmentation and compared them with other related methods, we trained our different proposed ensemble architectures on recent popular annotated medical imaging benchmarks as described in Section (4.3.1).

4.3.1 Datasets and Pre-processing

HVSMR-2016 [1] on Whole-Heart and Great Vessel Segmentation from cardiovascular magnetic resonance images (CMR) in congenital heart disease. Thirty training CMR scans from 10 patients are provided by the organizers of the HVSMR-workshop in MICCAI conference. Three different images have provided for each patient: a complete axial CMR image, the same image cropped around the heart and thoracic aorta, and a cropped short axis reconstruction. We evaluated the proposed cascade-GANs on this dataset where we utilized and trained each stage on different images from the same patients. We trained the first stage of cascade-GANs on the full axial volume from all patients. The second stage takes the cropped images around the heart and thoracic aorta as input, and the same image cropped short axis is served as the input of the third stage. Figure 4.8 shows three different images from HVSMR dataset.

Additional to HVSMR dataset, we applied the BraTS 2018 (described in Section 3.3.1), LiTS 2017 (reported in Section 2.2.2), and microscopic cell images

[1] http://segchd.csail.mit.edu/data.html

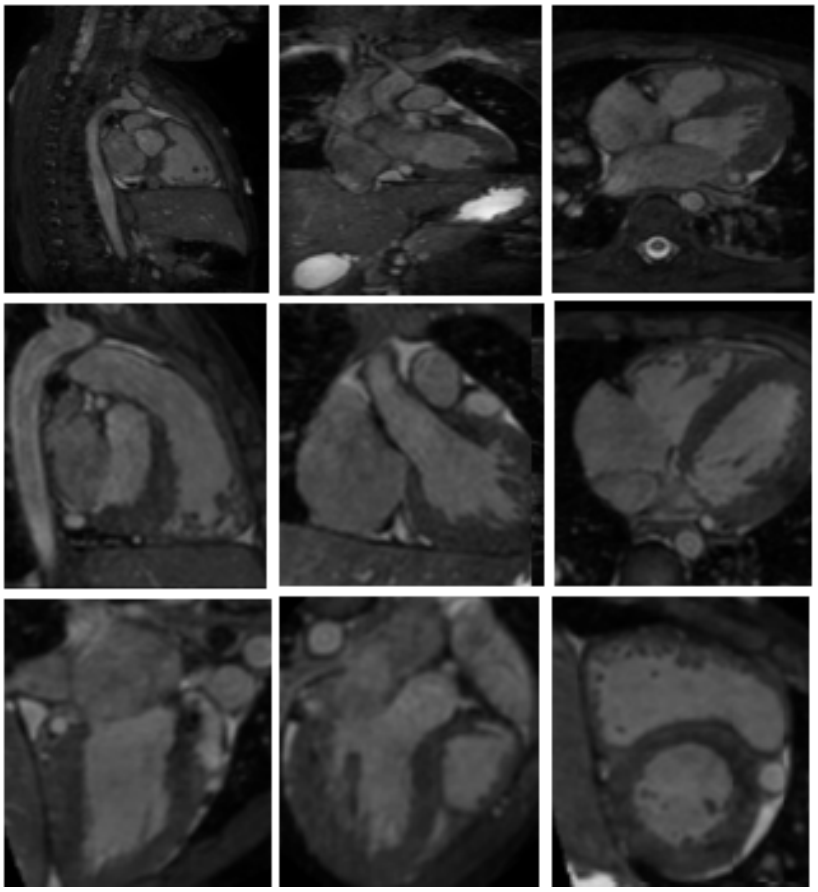


Figure 4.8: The axial, coronal, and sagittal view of images from HVSMR dataset where the first row shows complete axial CMR images, second and third row show the cropped around the heart and thoracic aorta and the cropped short axis respectively.

from ISBI 2015 (described in Section 2.2.2) which all explained in the previous chapters.

4.3.2 Implementation

Configuration

All the proposed ensemble methods are implemented based on a Keras library [59] with backend Tensorflow [62]. Our conditional refinement GAN code is publicly available [1]. We did not use any pre-trained model in our experiments and started training from scratch. All training and experiments are conducted on a workstation equipped with couple NVIDIA GPUs. in conditional refinement network, the learning rate is initially set to 0.0001. The RMSprop optimizer is used in the recurrent generator, discriminator, and refinement, it dividing the learning rate by an exponentially decaying average of squared gradients. We used Adadelta as an optimizer for cascade-GAN network that continues learning even when many updates have been done.

Our implemented cGAN [2], recurrent cGAN [3], and refinement model [4] are publicly available. At the recurrent architecture selected for both discriminator and generator is a bidirectional LSTM proposed by Graves et al. [97]. We used all 2D sequences from axial, coronal, and sagittal planes from both training and testing phases.

The cascade-GANs code is modified based on implemented conditional pix-to-pix and it is available on HPI-Deep Learning GitHub [5].

Conditional Refinement Architecture

In the conditional refinement GANs, a generator network is a modified UNet architecture with bidirectional LSTMs unit. The UNet architecture allows low-level features to shortcut across the network. The bidirectional LSTM provides inter as intra slice feature representation which is very important in sequential medical

[1] https://github.com/HPI-DeepLearning/MISS-GAN
[2] https://github.com/HPI-DeepLearning/MISS-GAN/tree/master/FirstStage
[3] https://github.com/HPI-DeepLearning/MISS-GAN/tree/master/FirstStage/Recurrent-cGAN
[4] https://github.com/HPI-DeepLearning/MISS-GAN/tree/master/SecondStage/GenerativeRefinement
[5] https://github.com/HPI-DeepLearning/SegMed

image analysis. The advantage of bidirectional LSTM appears when we connected features from down-convolution encoder with corresponding up-convolution decoder.

Our discriminator is fully convolutional Markovian PatchGAN classifier [40] which only penalizes structure at the scale of image patches. Unlike, the PatchGAN discriminator introduced by Isola et al. [40] which classified each N N patch for real or fake, we have achieved better results for a task of semantic segmentation in pixel level where we consider N=1. Moreover, since we have a sequential data, the bidirectional LSTM added after the last CNN layer in discriminator network. We used categorical cross entropy [65] as an adversarial loss with combination of $\ell 1$ loss in generator network. More detail about each architecture is described in Table 4.1 which "BN", "NL" and "DO" are short for batch normalization, non-linearity, and dropout layer respectively.

Regarding the high imbalanced datasets, minority pixels with lesion label are not trained as well as majority pixels with non-lesion label. Therefore, we designed boosting ensemble network named refinement to tackle this issue. The refinement network has the same architecture as our recurrent generator. The refinement network takes the predicted output from cGAN and medical images. The refinement network outputs two binary masks: false positive and false negative.

Cascade-GANs Architecture

We adjust our generator network architecture is auto-encoder similar to UNet, and discriminator is fully convolutional network. Let $Conv_k$ denote a convolution-BatchNorm-ReLU layer with k filters. $ConvDrop_k$ and $\oplus$ denote a Convolution-BatchNorm-Dropout-ReLU layer with a dropout rate of 50% and concatenation respectively. All convolutions are 4×4 spatial filters applied with stride 2.

Convolutions in the encoder, and in the discriminator, downsample by a factor of 2, whereas in the decoder they upsample by a factor of 2.

All leaky ReLUs adopted with slope 0.2. All other discriminators follow the same basic architecture, with depth size of $1X1$ for receptive field (see Table 4.2).

Ensemble-GANs

Table 4.1: Network architecture and hyper parameter for conditional refinement framework.

Operation	Kernel	Strides	Feature maps	BN	DO	NL
Generator/Segmentor						
Convolution	2×2	2×2	128	✓	✓	ReLU
Convolution	4×4	4×4	256	✓	✓	ReLU
$5 \times$ Convolution	8×8	8×8	5×512	✓	✗	ReLU
Discriminator						
Convolution	2×2	2×2	128	✓	✓	ReLU
Convolution	4×4	4×4	256	✓	✓	ReLU
Fully connected			$4 \times 4 \times 1024$	✗	✗	Tanh
Refinement						
Convolution	2×2	2×2	128	✓	✓	Leaky ReLU
Convolution	4×4	4×4	256	✓	✓	Leaky ReLU
Fully connected	8×8	8×8	5×512	✓	✗	Leaky ReLU
Batch size	1 (4 Modalities)					
Leaky ReLU slope	0.2					
Learning rate	0.0002					
Optimizer	RMSprop, $\beta = 0.9$					
BatchNorm	$\epsilon = 0.00001$, $\beta = 0.98$					

We designed generative adversarial ensemble discriminative networks consists single 3D generator encoder-encoder with multiple 3D fully convolutional discriminator networks with different losses.

The 3D boxes in generator as well as in discriminator (in Figure 4.7) shows CNN layer with its number of features size. Table 4.3 shows the detail of our architecture.

We update generator G twice and then update discriminator D during the learning iteration to balance the overall learning process of generator and discriminator. Noise is sampled element-wise from zero mean Gaussian having standard deviation of 0.001 during training. Standard deviation is changed to 1 and sampling is done in same manner as above, when we evaluate our algorithm. Based on observation, this change in standard deviation is useful to maintain proper level of diversity as we have very small-size data. To get better training of generator and discriminator in our model, batch normalization is used right after every convolutional layer.

4.3.3 Evaluation and Discussion

We validated and evaluated the performance of different proposed ensemble models using on quality metrics introduced by the challenges organizers. The BraTS challenge organizer provided the online judgment system, our reported result is based on challenge evaluation platform [1].

We also evaluated the performance of our approach on CT images for semantic segmentation of liver and lesion using the quality metrics introduced in the LiTS 2017 from grand challenges [133]. We follow the quantitative measurement by HVSMR2016, and microscopic cell segmentation.

4.3.3.1 Evaluation Results by conditional Refinement Network

Heterogeneous Brain Tumor Segmentation:

The segmentation of the brain tumor from medical images is highly interesting in surgical planning and treatment monitoring. The goal of segmentation as

[1] http://braintumorsegmentation.org/

Table 4.2: Network architecture and hyper parameter for cascade-GAN framework.

Operation	Kernel/Strides	Feature maps	BN	DO	NL
		Generator			
Convolution	3×3×3	128	✓	✓	Leaky ReLU
Convolution	3×3×3	256	✓	✓	Leaky ReLU
7 × Convolution	8 × 8	7 × 512	✓	✗	ReLU
		Discriminator			
Convolution	2 × 2	128	✓	✓	Leaky ReLU
Convolution	4 × 4	256	✓	✓	Leaky ReLU
Convolution	8 × 8	512	✓	✓	ReLU
Fully connected		4 ×1024	✗	✗	Softmax/Sigmo
Number of generators	3				
Number of discriminators	3				
Batch size	10				
Leaky ReLU slope	0.2				
Learning rate	0.0002				
Optimizer	Adam, $\beta_1 = 0.5, \beta_2 = 0.99$				
BatchNorm	$\epsilon = 0.00001$, $\beta = 0.98$				
bias initialization	0				

Table 4.3: Network architecture and hyper parameter for ensemble-GAN framework.

Operation	Kernel/Strides	Feature maps	BN	DO	NL	Shared?
		Generator				
4 ×Convolution	2 × 2	128	✓	✓	ReLU	✓
4 ×Convolution	4 × 4	256	✓	✓	ReLU	✓
10 × Convolution	8 × 8	7 × 512	✓	✗	ReLU	✓
		Discriminator				
Convolution	5×5×5	128	✓	✓	ReLU	✓
Convolution	8×8×8	256	✓	✓	ReLU	✓
Convolution	8×8×8	512	✓	✓	ReLU	✓
Fully connected		4×1024	✗	✗	Sigmoid	✓
Number of generators	1					
Number of discriminators	1,2,3					
Noise	random Uniform (-1,1)					
Batch size	1					
Leaky ReLU slope	0.2					
Learning rate	0.0002					
Optimizer	Adam, $\beta_1 = 0.5, \beta_2 = 0.99$					
BatchNorm	$\epsilon = 0.00001$, $\beta = 0.98$					
bias initialization	0					

described by organizer [2, 3, 4, 5] is to delineate different tumor structures such as active tumorous core (TC), enhanced tumorous (ET), and edema or whole tumorous (WT) region.

Figure (4.9) shows qualitative results of the cGAN network and refinement network in detail. Based on Figure (4.9), the result shows good relation to the ground truth for the segmentation after refinement network. The final output is refined through eliminating false negative pixels and adding the false positive pixels.

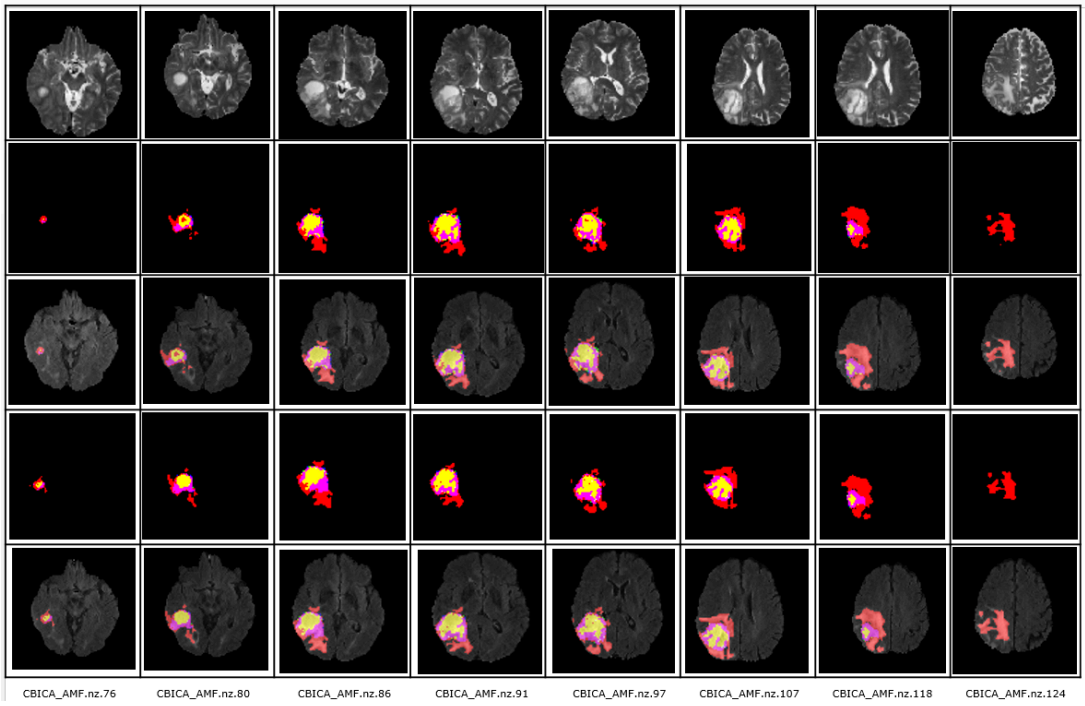


Figure 4.9: Visual results from our model on axial views of CBICA-AMF.nz.76-124 from the validation set. The first row shows Flair modality, while the second and fourth row shows the output results respectively from cGAN and refinement architecture. Third row shows the semantic segmentation masks from cGAN overlaid Flair modalities where the fifth row shows outputs after the refinement network. The red color codes the whole tumor (WT) region, while pink and yellow represent the enhanced tumor (ET) and the tumorous core (TC) respectively.

The Dice score, Hausdorff distance, sensitivity, and specificity are introduced by BraTS2017 as evaluation criteria for segmentation task. Table 4.4 and Table 4.5 present the brain segmentation results from proposed architecture and com-

pare them with other related methods based on the pre-proceeding report [134].

Table 4.4: Comparison of the achieved accuracy for semantic segmentation of different classes of tumor in terms of Dice and Hausdorff distance on validation data [2, 3, 4, 5] reported by the BraTS2017 organizer. The terms WT, ET, and TC are abbreviations of whole tumor region, enhanced tumor region, and core of tumor respectively.

Model	Dice			Hausdorff		
	WT	ET	TC	WT	ET	TC
RNN-cGAN+Refinement	0.86	0.64	0.73	7.22	8.30	11.04
cGAN	0.74	0.53	0.61	12.6	16.41	31.0
Recurrent-cGAN	0.79	0.60	0.68	11.73	14.54	25.83
3D-Ensemble-GANs	0.79	0.60	0.68	11.73	14.54	25.83
cascade-cGAN	0.79	0.60	0.68	11.73	14.54	25.83
Residual-Encoder [135]	0.82	0.62	0.57	-	-	-
Dilated-CNN [136]	0.36	0.77	0.34	2.23	3.3	5.4
FCN [137]	0.83	0.69	0.69	11.06	11.49	12.53
3D-Unet [138]	0.81	0.76	0.72	13.65	22.36	13.88
RNN [139]	0.84	0.71	0.73	4.6	4.18	8.18
Seq-3D-Unet [140]	0.76	0.90	0.84	13.33	8.9	14.1
Masked-Vnet [141]	0.86	0.71	0.63	5.43	8.34	11.17
3D-Seg-Net [142]	0.79	0.60	0.64	23.33	21.09	26.01
Nifty-Net [143]	0.83	0.71	0.68	27.49	17.35	31.34
3D-CNN [144]	0.82	0.46	0.56	9.56	13.8	14.7
biomedia [145]	0.90	0.73	0.79	4.2	4.5	6.5
UCL-TIG [108]	0.90	0.78	0.83	3.8	3.2	6.4
MIC-DKFZ [110]	0.89	0.73	0.79	6.9	4.5	9.4

From Table 4.4, the cGAN network (in second line) with one generator and discriminator achieved 12% less accuracy for whole tumor region segmentation compared to the segmentation results after the refinement network. In the first stage, the generator is trained by true positive and true negative masks. Meanwhile, the discriminator network tests how true is the predicted mask created by the generator. On the top of cGAN, the refinement learns the false negative and false positive masks. Table 4.5 presents discovery of false negative rate (1-recall) and false positive rate (1-specificity) in detail of network architecture. The final

Table 4.5: Comparison and the achieved accuracy for semantic segmentation in terms of false negative rate, false positive rate on validation set. WT, ET, and TC are abbreviations of whole tumor region, enhanced tumor region, and core of tumor respectively.

Model	False Negative Rate			False Positive Rate		
	WT	ET	TC	WT	ET	TC
RNN-cGAN+Refinement	0.11	0.16	0.29	0.02	0.02	0.02
cGAN	0.22	0.34	0.32	0.02	0.04	0.03
Recurrent-cGAN	0.19	0.32	0.30	0.02	0.03	0.02
3D-Ensemble-GANs	0.22	0.34	0.32	0.02	0.04	0.03
cascade-cGANs	0.19	0.32	0.30	0.02	0.03	0.02
Residual-Encoder [135]	0.83	0.66	0.59	.0.99	0.99	0.99
Dilated-CNN [136]	0.36	0.77	0.34	2.23	3.3	5.4
FCN [137]	-	-	-	-	-	-
3D-Unet [138]	-	-	-	-	-	-
RNN [139]	0.84	0.74	0.68	0.99	0.99	0.99
Seq-3D-Unet [140]	0.76	0.90	0.84	13.33	8.9	14.1
Masked-Vnet [141]	0.87	0.72	0.61	0.99	0.99	0.99
biomedia [145]	0.11	0.22	0.24	-	-	-
UCL-TIG [108]	0.09	0.23	0.18	-	-	-
MIC-DKFZ [110]	0.11	0.21	0.22	-	-	-

masks computed from the cGAN (or recurrent-cGAN) network with eliminating false negative and adding false positive predicted by the refinement network.

Regarding results of false discovery rate presented in Table 4.5, we have achieved good results as second and third-ranked teams in BraTS2017 competition when the segmented masks computed by recurrent conditional GAN and refinement network. Regarding quantitative results by Table 4.4 and Table 4.5, the networks substituted by LSTM unit predicted more accurate results.

Regarding the pre-proceeding by BraTS2017, most of the participants applied deep learning techniques, especially discriminative approaches to solve the brain tumor segmentation. Since most of the participants applied the discriminative models for segmentation, another key success of is adversarial training with the unique architecture consist of the couple discriminators and a generator network. Regarding first three rows obtained on Table 4.4 and Table 4.5, accuracy in terms of Dice, Hausdorff, and sensitivity is better when the both discriminator and generator network substituted by LSTM unit. Higher accuracy is expected due to well suit of our sequential training and testing architecture where the RNN gets the previous time steps image as context and the current time steps image.

The BraTS 2017 training dataset comprised 230 subjects with high-grade and 72 subjects with low-grade gliomas. There were 42 and, 114 subjects with mixed high- and low-grade gliomas respectively for validation and testing. Every participating group had two weeks to process their methods on the validation dataset and infinite time for submitting their segmentation results to the online evaluation system. In test time, every group had 48 hours from receiving the test subjects to process them and submit their segmentation results to the online evaluation system. The average value of the Dice coefficient is 0.85 in test time, which the results from Table 4.6 obtained and evaluated by challenge organizer. Since the results of the challenge in testing are not publicly available, we are not able to compare the performance of the different approaches in the test time. Table 4.6 shows the results from the validation set by the proposed method with patient-wise batch-norm and without Gaussian noise, evaluated at the BraTS2017 online judge system.

It is important to mention that our method takes only 58 seconds to segment one MR brain image consisting of 155 slices at a testing time.

Table 4.6: The achieved accuracy for brain tumor semantic segmentation by proposed conditional refinement GAN in terms of Dice, sensitivity, specificity, and Hausdorff distance reported by the BraTS-2017 organizer.

Evaluation	Validation			Test		
	WT	ET	TC	WT	ET	TC
Dice	0.86	0.64	0.73	0.85	0.61	0.72
Sens	0.89	0.84	0.71	-	-	-
Spec	0.98	0.98	0.97	-	-	-
Hdfd	7.22	8.30	11.04	8.73	59.2	25.9

Simultaneous Liver and Lesion(s) Segmentation:
Liver cancer is one of the most common types of cancers around the world [96] and CT images are widely used for diagnosis of hepatic diseases. The proposed method was trained on the public clinical CT dataset from LiTS2017 competition.

Figure 4.10 shows segmentation output in detail of conditional GAN in the left followed by refinement output in the right side of the figure.

In this competition, the primary metric is the Dice score. A volume overlap error (VOE), relative volume difference (RVD), average symmetric surface distance (ASSD), and maximum symmetric surface distance (MSSD) are considered for the evaluation of predicted region of liver and lesion(s). Table 4.7 and Table 4.8 describe the quantitative results and comparisons with top-ranked methods from LiTS leader-board [1].

To have the better understanding of the performance gains, we analyze the achieved accuracy on imbalanced liver tumor segmentation dataset where we can see unbalancing labels between large body organ and very small lesions. Based on the leader-board, most top-ranked models used cascade networks to segment simultaneously [146] or separately [71, 73] liver as well as lesion. The cascade networks provide good solution against imbalanced labeling. Unlike other cascade models [73, 146], our solution is based on cGAN approach where we segmented liver and lesions in semi-supervised manner. In second step, the refinement network trained with false positive and false negative masks which is important key

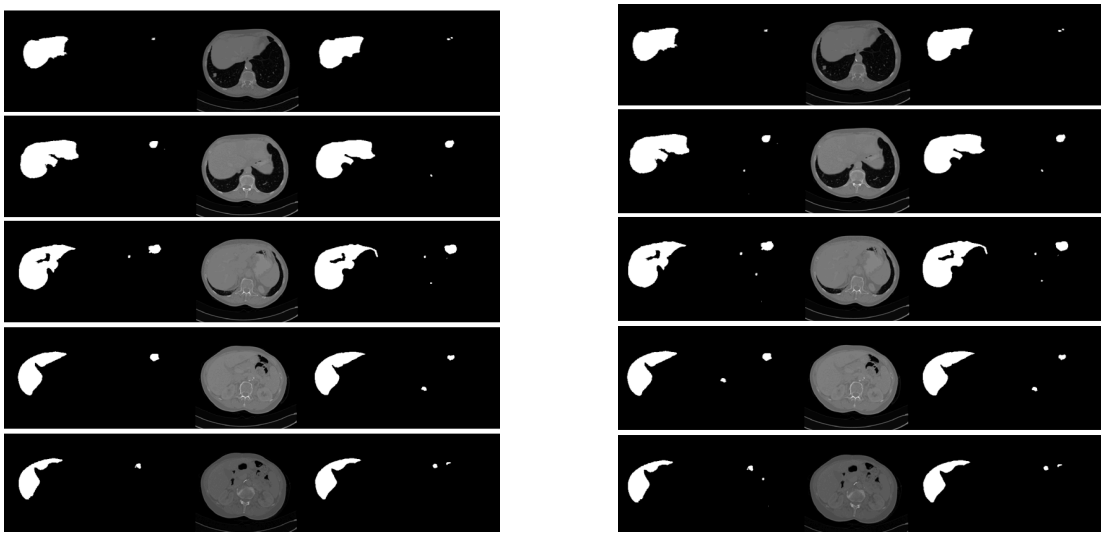
[1] https://competitions.codalab.org/competitions/

Table 4.7: The achieved accuracy for simultaneous liver and lesions segmentation in terms of Dice score and average surface distance on the test data where the 1 is the index of liver and 2 for lesions.

Approaches	Dice1	Dice2	ASD1	ASD2
cGAN+Refinement	0.94	0.83	1.4	1.6
cGAN	0.85	0.81	1.8	2.1
UNet	0.72	0.70	19.04	19.04
Cascaded-UNet [70]	0.93	0.93	2.3	2.3
VGG-FCN [71]	0.36	0.77	0.34	2.23
ResNet [71]	0.36	0.77	0.34	2.23
UNet+3DCRF [70]	0.95	0.50	0.92	1.3
ResNet+3DCRF [71]	-	-	0.84	13.33
ResNet+Fusion [71]	0.95	0.50	0.84	13.33
SuperAI	0.96	0.81	-	1.1
H-Dense+ UNet [146]	0.96	0.82	1.45	1.1
coupleFCN [73]	0.78	0.77	-	-

Table 4.8: The top two rows show achieved accuracy for the simultaneous liver and lesions segmentation in terms of Dice score and average surface distance on the test data.

Architecture	VOE	RVD	ASD	MSD
cGAN+Refinement	14	-6	6.4	40.1
cGAN	21	-1	10.8	87.1
Cascaded-UNet [70]	22	-3	9.5	165.7
ResNet+Fusion [71]	16	-6	5.3	48.3
SuperAI	36	4.27	1.1	6.2
H-Dense+ UNet [146]	39	7.8	1.1	7.0
coupleFCN [73]	35	12	1.0	7.0



(a) (b)

Figure 4.10: Segmentation results obtained by cGAN (a) compared to the refinement output (b). In each subfigure, the first two left columns show the ground truth manual segmentation of the liver and lesion(s). The two last right columns from (a,b) show the predicted liver and lesion(s) at the first and second stages.

aspect for clinical routine application. The final semantic masks computed by adding false positive pixels and subtracting false negative pixels.

Table 4.7 describes our obtained result for liver segmentation and lesions in terms of the Dice score 0.94 and 0.83 respectively. Based on Table 4.7 and with comparison of the first two rows, we can see the effect of refinement network on final results which has increased up to 9% for liver segmentation and similarly up to 2% for the lesions segmentation.

In the LiTS dataset, lesions with an approximate diameter equal to or larger than 10 mm were defined as a large one, while a small lesion has a diameter of less than 10 mm. Our method achieved an average Dice of 0.90 and ASD of 1.6 in lesion segmentation which obviously, can distinguish small and large lesions. We provided more qualitative results when the segmentation target on lesions are small or large size.

In addition, our algorithms are very fast, and it takes only 100 seconds for the simultaneous segmentation of liver and lesion from CT images with 280 slices,

each sized 512 x 512. The complex and heterogeneous structures of the predicted liver and all lesions from local test set are depicted in Figure. 4.10.

3.3.3 Microscopic Cell Segmentation:

Microscopy cell images are key component of the biological research process and automatic cell segmentation is helpful application for clinical routine. We evaluated our method on microscopic cell from human breast carcinoma (MDA231). MDA231 consists of 96 images with segmented ground truth files by experts.

Figures 4.11 and 4.13 compare the qualitative results from test set when the network was trained with and without patient-wise mini-batch normalization. The patient-wise mini-batch normalization provided normalization for any layer of neural network based on all available 2D images from same patient. Through patient-wise normalization technique, we normalized the activation of the previous layer for each patient batch.

Based on qualitative results and Figure 4.11, our network is able to learn from few samples (MDA231) as well as large sample dataset (BraTS2017). We compared quantitative results with the state-of-the-art segmentation method. The quantitative results of individual cell segmentation are detailed in Table (4.9). Obviously, we can see that diversity and the number of images did not have a major effect on the final result.

As shown in Figure 4.12 and Table 4.9 the Gaussian noise negatively influence the segmentation results especially when the trained dataset has few samples. We had same policy for data augmentation on all datasets. We explored during raining the large dataset, when the generator network takes Gaussian noise vector besides medical images, act mostly same as without noise vector and there is minimum differences in the output samples. In contrast, trained network with few samples along with noise vector has negative effect on the final outputs.

4.3.3.2 Evaluation Results by cascade-GANs

The proposed cascade-GANs is trained on 80% training data released by the HVSMR-2016 benchmark, which consists of 24 CMR images. We used all provided images (full, axial crop, and axial short) from three axes of x, y, and z

Figure 4.11: Microscopic cell segmentation results obtained by cGAN+Refinement network with patient-wise mini-batch normalization and without Gaussian noise.

Table 4.9: The achieved accuracy for cell segmentation in terms of intersection over union on the MDA231 data

Approaches	Dice	Spec	Sen	FPR	FNR
cGAN+Refinement	0.93	0.93	0.92	0.07	0.08
RNN-GAN	0.91	0.90	0.91	0.10	0.09
cGAN	0.90	0.89	0.91	0.11	0.09
UNet [66]	0.92	-	-	-	
KTH-SE [67]	0.79	-	-	-	
MSER [68]	0.75	-	-	-	
Greedy [69]	0.85	-	-	-	

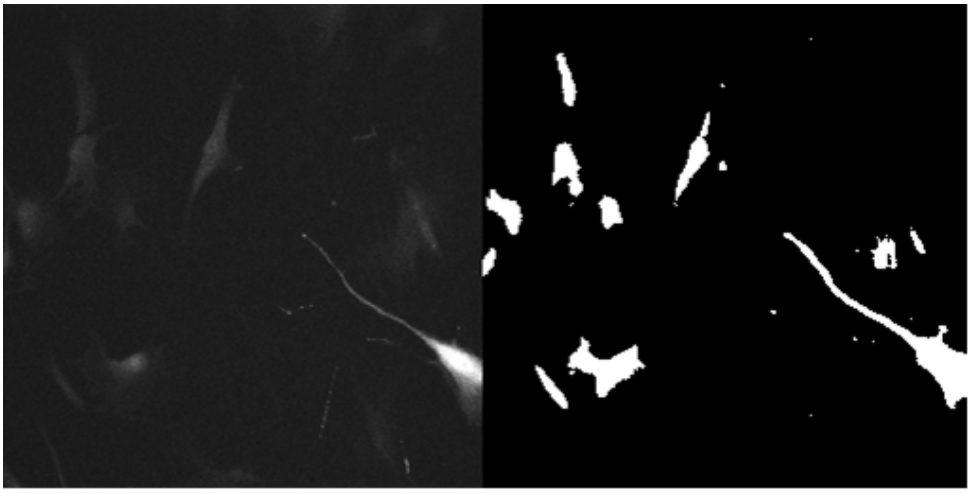


Figure 4.12: Microscopic cell segmentation results obtained by cGAN when the cGAN model trained with additional Gaussian noise as input.

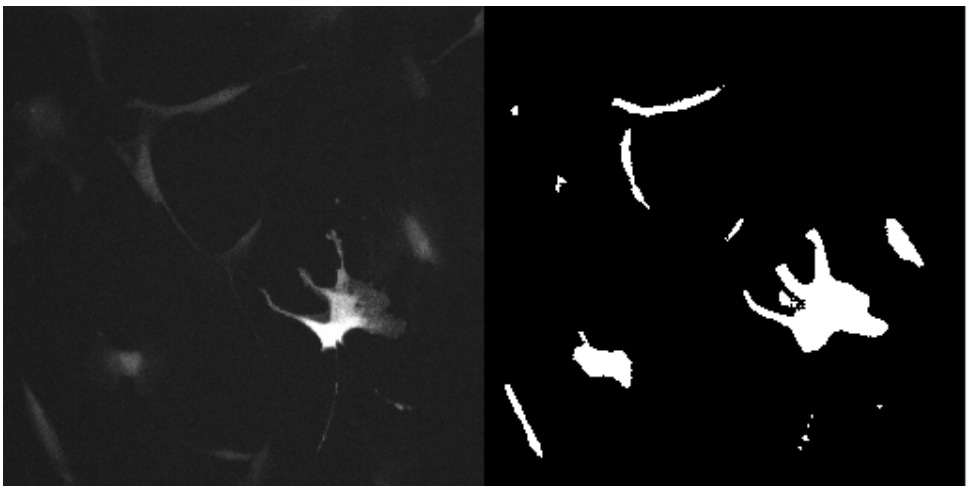


Figure 4.13: Microscopic cell segmentation results obtained by cGAN without patient-wise mini-batch normalization.

for training and testing. From Table 4.10, we can also infer that the cascade of GANs can improve training accuracy in the term of Dice by up to 8%.

Qualitative results are shown in Figure 4.14. The training takes around three days for a total of 100 epochs on parallel Pascal Titan X GPUs for semantic segmentation tasks. We train three GANs iteratively where each GAN is trained not only with the feature data, but also with the probability map obtained from the previous GANs, which gives to the GAN additional context information. At

Table 4.10: The evaluation result of the semantic segmentation network by cGAN compared to cascade-GANs. The first top rows demonstrates the performance gains by using cascade of cGANs. We compared our results with the best performance team reported by the HVSMR-2016. Performance of our method cascade-GAN on the testing datasets in terms of average distance of boundaries (Adb) and Dice. Label 1 indicates the myocardium tissue while label 2 stands for the blood pool.

Method	Adb1	Adb2	Dice1	Dice2	Sen1	Sen2	Spec1	Spec2
cascade-GAN	1.02	0.87	0.80	0.93	0.87	0.90	0.96	0.99
cGAN	1.19	1.07	0.72	0.89	0.82	0.88	0.94	0.99
UNet(2D)	2.04	1.82	0.68	0.81	0.78	0.74	0.91	0.99
Shahzad et al. [147]	1.10	1.55	0.75	0.89	-	-	-	-
Yu et al. [148]	0.99	0.86	0.78	0.93	-	-	-	-
Wolterink et al. [149]	0.89	0.96	0.80	0.93	-	-	-	-

testing time, the features will be processed for each GAN one after the other, concatenating the probability map to the input features. The proposed network provides two predicted masks of blood pool and myocardium for each 2D image in less than 30 ms.

The results show good relation to the ground truth for the blood pool. The average value of the Dice index is around 0.93, which is the same as the result of the HVSMR-2016 challenge winner [148]. The main source of error here is the inability of the method to completely segment all the great vessels where the average Dice score is 0.80. Higher accuracy is expected while each stage trains on shared extracted features and weights from the previous stage.

In order to compare the performance of cascade-GANs with other ensemble approaches, we train the network for 120 epochs that 60, 40, and 20 as first, second, and third stage. The quantitative results are shown in Figure 4.15.

4.3.3.3 Evaluation Results by Ensemble-GANs

For evaluation of ensemble-GANs, we applied BraTS 2018 benchmark to compare it with other proposed ensemble framework and related approaches. We compare different implementations of the proposed ensemble-GANs architecture and also evaluate the effectiveness of different losses on handling imbalanced data for task

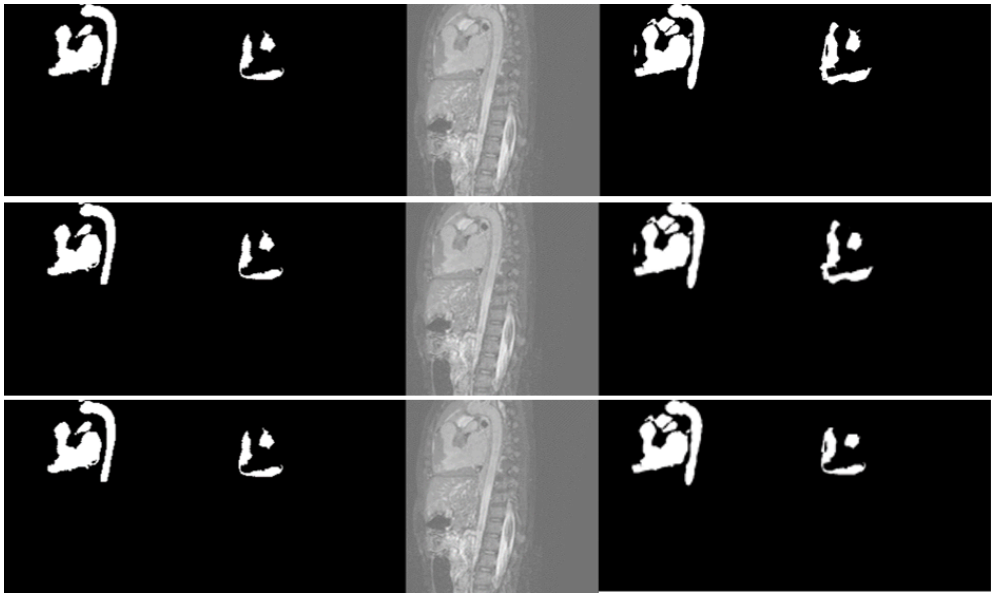


Figure 4.14: The visualization results from three stages: The first two columns show the ground truth annotated by medical experts from the HVSMR2016, and the third column shows the z plan of CMR data, which is the input of context-aware cGAN. The fourth and fifth column show the predicted results by cascade-cGAN in different stages. The first row shows the output from the first stage after 50 epochs; the second and third rows are the output after 75 and 100 epochs from the second and third stages.

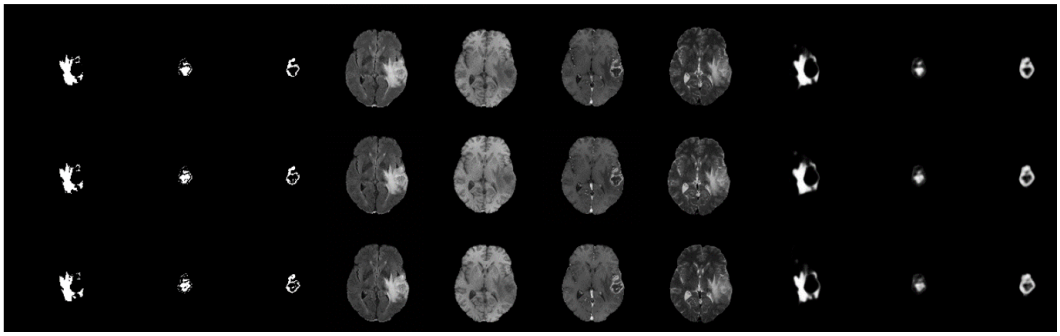


Figure 4.15: Brain tumor semantic segmentation by cascade-GAN. The first three columns show the ground truth mask, columns 4-7 show multi-modal MR images as input for cascade architecture while columns 8-10 are the predicted results.

of semantic segmentation. Specifically, we compare the following implementations:

- $3D-G, 3D-D_1$. The GAN framework consists a 3D hourglass as generative

model and a 3D FCN as discriminator network where the discriminator has $\ell 1$ as loss function and binary cross-entropy is used as an adversarial loss.

- $3D-G, 3D-D_2$. The GAN framework consists a two 3D UNet as generator and a 3D FCN as discriminator network where the discriminator has Dice as loss function and binary cross-entropy is used for adversarial loss.
- $3D-G, 3D-D_3$. An hourglass network including two 3D UNet is trained as a generator model and a 3D FCN as discriminator network where the discriminator has categorical accuracy as loss function and binary cross-entropy is used as an adversarial loss.
- $3D-G, 3D-D_{1,2}$ The generative architecture is similar to previous model while discriminative is 3D FCN. More specifically, each discriminator is trained to minimize the average negative cross-entropy between predicted and the actual labels in different distance. We minimize error regarding Dice and $\ell 1$.
- $3D-G, 3D-D_{1,3}$. Similar to previous model the GAN network trained with binary cross-entropy, Dice, and categorical accuracy.
- $3D-G, 3D-D_{1,2,3}$. We keep the network architecture similar previous model and this time the GAN framework is substituted with binary cross-entropy, Dice, categorical accuracy and $\ell 1$.

All described above each framework is trained up to 100 epochs, Table 4.3 gives more detail about architecture. The generator takes four different MRI modalities provided by BraTS 2018, the network has fixed size of $128 \times 128 \times 32$ voxels and a spatial resolution of $1 \times 1 \times 1.5$ millimeters. During every training iteration, we fed as input to the network randomly cropped, rotation [-10, 10] of the training images through a $2 \times 2 \times 2$ grid of control-points.

The quantitative results for brain tumor semantic segmentation reported on Table 4.12 and Table 4.11 by several ensemble GAN architecture. The networks were trained with different number of discriminator and various losses. The better performance for handling imbalanced data achieved when the generator minimize the semantic segmentation error by three different discriminator (see Table 4.12-fifth row) rather than in cascade architecture (see Table 4.12-seventh row).

Table 4.11: Comparison of the achieved accuracy for semantic segmentation of different classes of tumor in terms of Dice and Hausdorff distance on subset of validation data reported by the BraTS 2018 organizer. WT, ET, and TC are abbreviations of whole tumor region, enhanced tumor region, and core of tumor respectively.

Model	Dice			Hausdorff		
	WT	ET	TC	WT	ET	TC
$3D-G, 3D-D_1$	0.84	0.64	0.71	7.22	12.30	8.04
$3D-G, 3D-D_2$	0.84	0.67	0.72	7.6	11.41	8.12
$3D-G, 3D-D_3$	0.83	0.63	0.71	7.73	11.65	7.83
$3D-G, 3D-D_{1,2}$	0.87	0.65	0.79	6.11	10.54	7.91
$3D-G, 3D-D_{1,3}$	0.89	0.65	0.81	6.23	9.22	6.89
$3D-G, 3D-D_{1,2,3}$	0.88	0.70	0.81	6.73	9.28	7.11
cascade-cGAN	0.87	0.60	0.68	11.73	8.54	7.43
Refinement	0.88	0.73	0.82	6.67	9.28	6.94

Table 4.12: Comparison of the achieved accuracy for semantic segmentation of different classes of tumor in terms of precision and sensitivity on subset of validation data reported by the BraTS 2018 organizer. WT, ET, and TC are abbreviations of whole tumor region, enhanced tumor region, and core of tumor respectively.

Model	Precision			Sensitivity		
	WT	ET	TC	WT	ET	TC
$3D-G, 3D-D_1$	0.88	0.66	0.79	0.82	0.63	0.74
$3D-G, 3D-D_2$	0.85	0.60	0.81	0.86	0.68	0.72
$3D-G, 3D-D_3$	0.84	0.67	0.80	0.80	0.74	0.61
$3D-G, 3D-D_{1,2}$	0.89	0.71	0.79	0.84	0.72	0.83
$3D-G, 3D-D_{1,3}$	0.88	0.69	0.81	0.89	0.78	0.79
$3D-G, 3D-D_{1,2,3}$	0.89	0.74	0.80	0.88	0.79	0.84
cascade-cGAN	0.88	0.66	0.79	0.82	0.63	0.74
Refinement	0.91	0.74	0.81	0.88	0.87	0.74

4.3.4 Discussion

Based on represented results, ensemble model is a proven model for improving the accuracy and better trade-off between precision and recall which able to mitigating imbalanced data. Here, we briefly count some advantage and disadvantage of each proposed ensemble framework: Ensembling makes the model more robust and stable thus ensuring decent performance on the test cases in most scenarios. The prediction result from aggregation of multiple classifiers is less noisy compared to individual. However, ensembling reduces the model interpretability and makes it very difficult to draw any crucial business insights at the end. It is time-consuming and thus might not be the best idea for real-time applications. The selection of models for creating an ensemble is an art which is really hard to master.

4.4 Summary and Extension

In this chapter, we introduced three different deep ensemble models namely 1) conditional refinement GAN, 2)cascade-GANs, 3)ensemble-GANs to address the imbalanced class problem in task of semantic segmentation.

The proposed conditional generative refinement network, consists of three components: a generative, a discriminative, and refinement networks ensemble learning. The generative network learns to the segment at the pixel level by getting feedback from the discriminative network according to the true positive and true negative maps. On the other hand, the refinement network learns to predict the false positive and the false negative masks produced by the generative model that has significant value, especially in medical application. The final semantic segmentation masks are then composed by the output of the three networks. The network is evaluated and tested on three recent medical imaging benchmark such as BraTS 2017, LiTS 2017, and microscopic cell 2015. The results show the the benefit of boosting ensemble for handling imbalanced data for semantic segmentation.

In this chapter, we introduced cascade-GANs made by three individual conditional GAN framework. These three stages are designed to share convolutional features and weights, where the later stages use the shared convolution

features from the previous stage and transfer the learned convolutional features and weights to the next stage. Each stage involves a cGAN with an adversarial loss and individual parameters. The final semantic segmentation computed from last stage. We tested the proposed approach on BraTS 2017, and HVSMR 2016. The predicted results by last stage shows the advantage of bucket of ensemble models.

The ensemble GAN consists by single generator and many discriminator proposed to mitigate the imbalanced data problem through simple averaging of ensemble expert classifiers. Our approach made the single generator not to constrain the output of a single discriminator, but, instead, to learn a dynamic ensemble of various discriminators. The model is evaluated by considering different losses on BraTS 2017 for semantic segmentation.

In future we plan to examine impact of building ensemble Bayesian model for handling imbalanced class problem. We plan to study effect of ensemble of discriminator's decision from different deep neural network architectures and with similar or different losses on single generator network.

However, the proposed ensemble approaches in this chapter are not limited for semantic segmentation task; these models have potential to apply for other imbalanced class distribution domain including fraud detection, few-shot learning, and autonomous driving.

Chapter 5

Learning Imbalanced Semantic Segmentation by Deep Mutual GANs

Learning representation is one of canonical objective for approaching most of the deep learning models. We address learning deep representation from imbalanced data by Mutual Information (MI) between independent ensemble classifiers.

This chapter investigates learning representation of imbalanced data by maximizing the mutual information between the ensemble of networks named Deep Mutual Generative Adversarial Networks (DM-GAN). The proposed deep mutual GAN framework consists of multiple small generators and multiple large discriminators. During the training process, each generator gets feedback from pre-trained discriminator network in adversarial way. Additionally, each generator learns to collaborate with other generators during the training process. Here, we explore the mutual information shared between independent generators helpful to mitigate the imbalanced class problem which is one of major differences between this chapter and chapter 4.

Similar to model distillation, we aim to transfer knowledge from a powerful network (teacher) as discriminator to a small network (student), generator, in order to meet high performance and mitigate imbalanced class problem. Our deep mutual GAN differs from the one-way transfer between a static pre-defined

discriminator and a generator in model distillation. Here a generator network has two losses: (1) adversarial loss to learn from powerful discriminator, (2) mutual information loss to learn from an ensemble of generators (each generators class posterior with the class probabilities of other generators). Trained in this way, it turns out that each generator in such a peer-teaching based scenario learns significantly better than when learning alone in a conventional supervised learning scenario.

Experimental results show that a designed GAN based framework benefits from mutual learning and achieve compelling results on semantic segmentation tasks. Importantly, it shown that mutual loss of a collection of simple generator networks handle imbalanced class problem.

5.1 Technical Background

Mutual Information

From probability and information theory, the Mutual Information (MI) is a measure of the mutual dependence between two variables. Accurately, MI quantifies the *amount of information* (in units such as Shannons, commonly called bits) obtained about one variable by observing the other variable. Mutual information can be equivalently expressed as:

$$I(X;Y) = H(X) - H(X|Y) \tag{5.1}$$

To understand Eq.(5.1), we first introduce two other quantities: the conditional entropy $H(X|Y)$ and the marginal entropy $H(X)$. Conditional entropy is a measure of the uncertainty in one variable after observing another. The conditional entropy of the random variable Y conditioned on the random variable X is defined as:

$$H(Y|X) = -\sum_{x\in X, y\in Y} p(x,y) log \frac{p(x,y)}{p(x)} \tag{5.2}$$

Thus, it is the average entropy of a variable after conditioning on knowledge of another variable.

Relative entropy, or the KullbackLeibler (KL) divergence, is a measure of how two probability distributions are different. The KL between two probability mass functions $p(x)$ and $q(x)$ over the same space X is defined as:

$$D_{KL}(p\|q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)} \tag{5.3}$$

In other words, it is the expectation of the logarithmic difference between the probabilities p and q, where the expectation is taken using the probabilities p. The KullbackLeibler divergence is defined only if for all $x, q(x) = 0$ implies $p(x) = 0$ (absolute continuity). Whenever $p(x)$ is zero the contribution of the corresponding term is interpreted as zero because: $\lim_{x \to 0} x log(x) = 0$

5.2 Deep Mutual GAN

5.2.1 Formulation

We formulate the proposed Deep Mutual GAN (DM-GAN, see Figure. 5.1) to study the effect of mutual information between untrained generators on handling the imbalanced class problem. During training, each powerful pre-trained discriminators send feedback with an individual loss to one generator through adversarial training. Meanwhile the generator collaborates and shares mutual information with other generators during training. We explore the mutual information loss together with adversarial loss useful for mitigating imbalanced data. This chapter builds upon Deep Mutual Learning (DML) introduced by Zhang et al. [150] and our previous 3DJointGAN [151].

However, there are several major differences between our proposed deep mutual GAN and the DML that make DM-GAN framework more suitable and effective for the task of image segmentation and handling imbalanced class problem. In contrast to the DML approach, in our framework, each discriminator substituted by different losses to mitigate the imbalanced class problem and minimize the generator error on a different aspect of quality.

Firstly, we describe the DM-GAN architecture consist of two generators G and two pre-trained powerful discriminators D. Secondly, we explain and examine our proposed DM-GAN by two different implemented architecture (see Section Section 5.2.2) and the extension to more networks defined in Section 5.2.3).

Assume the proposed DM-GAN framework contains couple of G and D networks. Each G learns mapping from multimodal MRI images $X = x_i$, and Gaussian vector z to desired semantic segmentation mask $Y = y_i$; $G : \{x_i, z\} \rightarrow \{y_i\}$ where each pixel belong to $1, 2, ..., M$ classes. Meanwhile, D classifies whether a generated segmentation mask by G, in pixel or voxel level belong to right class correct and during training tries to minimize the G's error.

The generators have similar architecture but each one gets feedback from specific discriminator losses. The discriminators have similar architecture and substituted with different losses and communicate only with one generator. The objective for one generator (See Eq. 5.4) and one discriminator (See Eq. 5.5) is similar:

$$\mathcal{L}_G = E_{x,z}[log(1 - D(x, G(x, z)))] \tag{5.4}$$

$$\mathcal{L}_D = E_x logD(x) - log(1 - D(x, G(x, z))) \tag{5.5}$$

We aim to mitigate the imbalanced class problem by approaching mutual information loss between predicted outcome by each generator; we followed the same mutual loss is defined by Zhang et al. [152]:

$$\mathcal{D}_{KL}(p_2||p_1) = \sum_{i=1}^{N}\sum_{m=1}^{M} p_2(x_i) log \frac{p_2^m(x_i)}{p_1^m(x_i)} \tag{5.6}$$

where the probability of class m for sample x_i given by deep generator neural network is computed as:

$$p_1^m(x_i) = \frac{exp(z^m))}{\sum_{m=1}^{M} exp(z^m)} \tag{5.7}$$

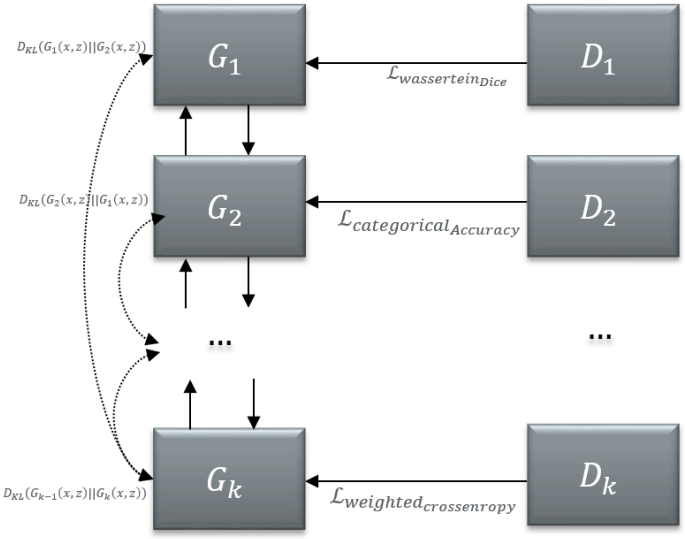


Figure 5.1: Deep Mutual GAN (DM-GAN), where each G network is trained with a supervised learning loss from individual discriminator, and a Kullback Leibler Divergence to match the probability estimates of its generators.

where the logit z is the output of the softmax layer in $G1$. Therefore, the final loss function is calculated by:

$$\mathcal{L}_{G_1} = \mathcal{L}_{G_{D_1}} + \mathcal{D}_{KL}(p_2||p_1) \tag{5.8}$$

$$\mathcal{L}_{G_2} = \mathcal{L}_{G_D 2} + \mathcal{D}_{KL}(p_1||p_2) \tag{5.9}$$

During training in this way, each generator network learns both to correctly predict the actual label of training instances by specific discriminator loss as well as to match the probability estimate of its peer (KL mimicry loss). Our work differs from Zhang et al. [152] by different architecture and different discriminator losses. Here, we implemented two different DM-GAN in 3D (Section 5.2.2.1) and 2D (Section 5.2.2.2) medical image semantic segmentation.

5.2.2 Network Architecture

As mentioned before, we consider two different GAN based architecture to see the impact of mutual learning for mitigating imbalanced data.

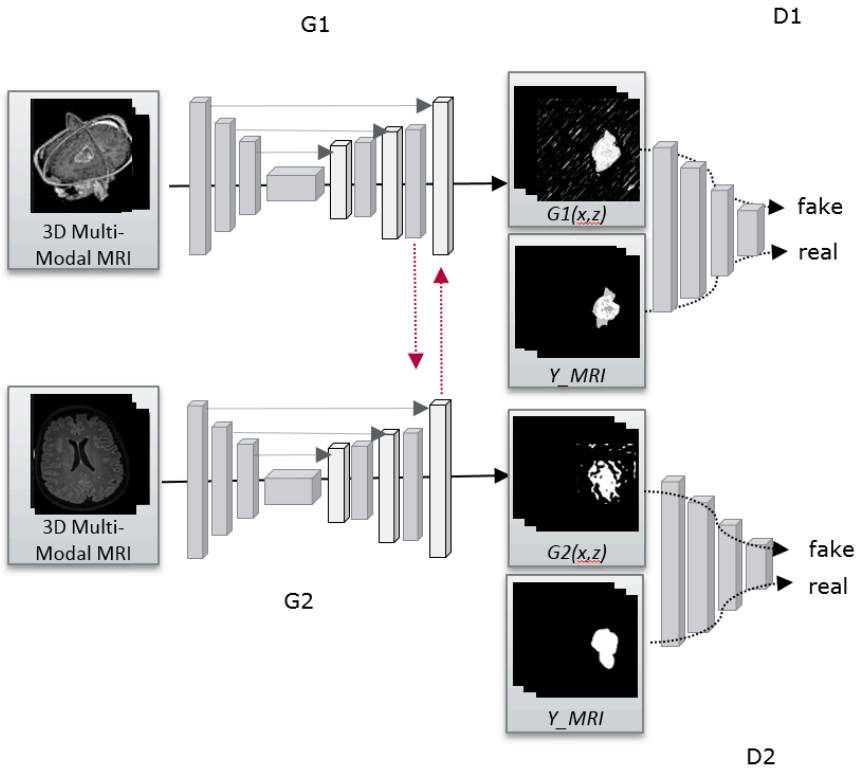


Figure 5.2: 3D Deep Mutual GAN (DM-GAN), conposed by couple GAN framework and each G net is trained with a supervised learning loss from individual discriminator, and a Kullback Leibler Divergence.

5.2.2.1 3D Deep Mutual Generative Adversarial Network for Semantic Segmentation

3D DM-GAN as illustrated in Figure 5.2 is designed for learning a imbalanced class problem in task of semantic segmentation. Similar to Zhang et al. [152], the models are learned with the same mini-batches. At each iteration, we compute the predictions of the two models and update both networks parameters according to the predictions of the other. Here, the generators start learning from scratch while the discriminators are pretrained 3D FCN models with different losses

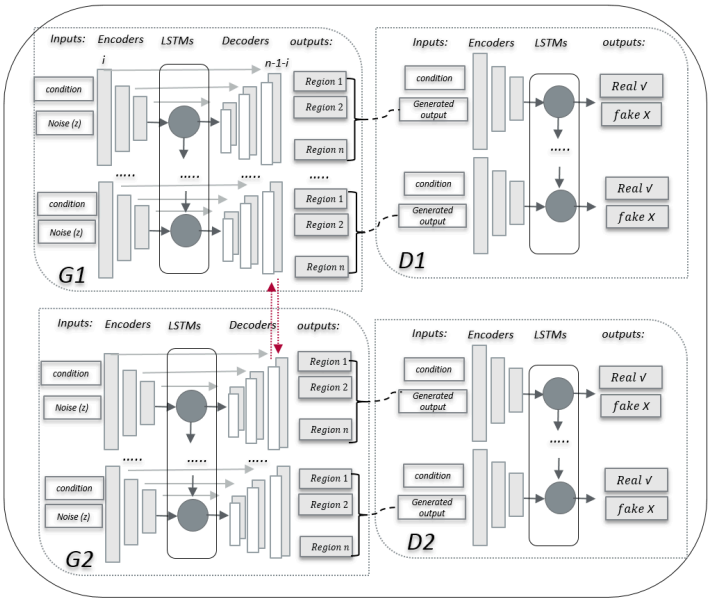


Figure 5.3: 2D recurrent DM-GAN consists of couple recurrent-GAN frameworks.

5.2.2.2 2D Deep Mutual Generative Adversarial Networks for Semantic Segmentation

As shown in Figure 5.3, in the recurrent setting, the couple of frameworks take 2D sequence of MRI or CT images and outputs are 2D sequence of semantic segmentation. Both segmentor and discriminator substitutes with bidirectional LSTM unit in a bottleneck. Here, the G is a fully convolutional encoder-decoder network with a bidirectional LSTM unit in bottleneck. The D is a fully convolutional encoder substitutes with bidirectional LSTM and classifier between the original pixel's label from ground truth and synthesized pixel value created by generator. In this way, each generator communicated and shared mutual information between themselves.

5.2.3 Extension

The proposed DM-GAN approach naturally extends to more networks. Assume K networks $g_1, g_2, ..., g_k$, $(K \geq 2)$, the objective function for optimising g_k, $(1 \leq$

$k \leq K$) is following:

$$\mathcal{L}_{G_k} = \mathcal{L}_{adv}(D_k, G_k) + \frac{1}{k-1}\sum_{k=1}^{K} \mathcal{D}_{KL}(p_1||p_k) \tag{5.10}$$

where the k generators collaborate with K networks. In this way, DM-GAN for each generator takes the other $K1$ networks in the cohort as teachers to provide mimicry targets. Equation 5.8 is now a special case of 5.10 with K = 2.

We trained DM-GAN by considering $K = 2$ since the optimization of DM-GAN with more than two networks need distributed learning strategy. The distributed model, each network learn on one device and passing the small probability vectors between devices. DML model [152] shown having more than two student network (generator) leads better performance compared to distillation approach because the model averaging step to building the teacher (pretrained discriminator) ensemble makes the teachers posterior probabilities more peaked at the true class, thus reducing the posterior entropy over all classes. It is, therefore, contradictory to one of the objectives of DML, which is to produce robust solutions with high posterior entropy. The future direction of this chapter is implementing 2D and 3D deep mutual GAN architecture consisting more than two generators where the generator learns and communicate in a distributed way.

5.3 Experiments

We validated the performance of proposed DM-GAN for semantic segmentation of imbalanced medical imaging on two implemented architectures. We tested two architectures on real patient data obtained from the BraTS 2018 [2, 3, 4, 5] for brain tumor semantic segmentation.

5.3.1 Datasets and pre-processing

The BraTS 2018 benchmark [2, 3, 4, 5] prepared 1,140 MR images in multi modal scans, NIFTI format on (a) native (T1) and (b) post-contrast T1-weighted (T1Gd), (c) T2-weighted (T2), and (d) T2 Fluid Attenuated Inversion Recovery

(FLAIR) volumes, and were acquired with different clinical protocols and various scanners from multiple (n=19) institutions.

All the imaging datasets have been segmented manually, by one to four raters, following the same annotation protocol, and their annotations were approved by experienced neuro-radiologists. Annotations comprise the GD-enhancing tumor (ET label 4), the peritumoral edema (ED label 2), and the necrotic and non-enhancing tumor core (NCR-NET label 1), as described in the BraTS reference paper [2]. The provided data are distributed after their pre-processing, i.e., co-registered to the same anatomical template, interpolated to the same resolution and skull-stripped by organizer.

BraTS 2018 provided 420 MR images in four modalities without annotated file for validation time. Participate in challenge are able to submit many times the result on online platform. we validated our proposed approach on validation set using the 2018 online platform of BraTS challenge.

To prevent over-fitting, we added data augmentation to each dataset such as randomly cropped, re-sizing, scaling, rotation between -10 and 10 degrees, and Gaussian noise applied on training and testing time for both datasets.

5.3.2 Implementation

This section introduces the detail of configuration and implemented 3D and 2D DM-GAN architecture.

Configuration of 3D Deep Mutual-GAN

The 3D DM-GAN is implemented based on a Keras library [59] with back-end Tensorflow [62] supporting 3D convolutional network. All training and experiments were conducted on a workstation equipped with a multiple GPUs. The detail of architecture is shown in Table 5.1.

The learning rate was initially set to 0.0002. The Adam optimizer was used in both the generator and the discriminator that continues learning even when many updates have been done. The model is trained for up to 150 epochs.

Configuration of 2D recurrent Deep Mutual-GAN

Table 5.1: Network architecture and hyper parameter for 3D Deep Mutual GAN framework.

Operation	Kernel/Strides	Feature maps	BN	DO	NL
		Generator			
Convolution	2×2×2	128	✓	✓	ReLU
Convolution	4×4×4	256	✓	✓	ReLU
3× Convolution	8×8×8	3×512	✓	✗	ReLU
		Discriminator			
Convolution	5×5×5	128	✓	✓	ReLU
2×Convolution	8×8×8	256	✓	✓	Leaky ReLU
4×Convolution	8×8×8	512	✓	✓	Leaky ReLU
Fully connected		4× 1024	✗	✗	CAcc /Softmax
Number of generators	2				
Number of discriminators	2				
Noise	random Uniform [-1,1]				
Batch size	2				
Leaky ReLU slope	0.2				
Learning rate	0.0002				
Optimizer	Adam, $\beta_1 = 0.5, \beta_2 = 0.99$				
BatchNorm	$\epsilon = 0.00001$, $\beta = 0.98$				
P_1	softmax(logits1)				
P_2	softmax(logits2)				
bias initialization	0				

The 2D DM-GAN is the extended version of our recurrent-GAN [61], implemented based on a Keras library [59] with back-end Tensorflow [62]. Similar to 3D version, all training and experiments were conducted on a workstation equipped with a multiple GPUs. The detail of architecture is shown in Table 5.2. The both discriminator networks has same architecture and pretrained, during training we fine tune weights regarding two new losses: a wasserstein Dice for first discriminator and weighted cross-entropy for second discriminator. The training for generator start from scratch. In this work, the recurrent architecture selected for both discriminator and generator is a bidirectional LSTM [97].

Table 5.2: Network architecture and hyper parameter for 2D recurrent Deep Mutual GAN framework.

Operation	Kernel/Strides	Feature maps	BN	DO	NL
		Generator			
Convolution	3×3	128	✓	✓	ReLU
Convolution	5×5	256	✓	✓	ReLU
3× Convolution	8×8	3×512	✓	✗	ReLU
		Discriminator			
Convolution	5×5	128	✓	✓	ReLU
2×Convolution	8×8	256	✓	✓	Leaky ReLU
4×Convolution	8×8	512	✓	✓	Leaky ReLU
Fully connected		4× 1024	✗	✗	WDice /Wcros
Number of generators	2				
Number of discriminators	2				
Noise	random Uniform [-1,1]				
Batch size	2				
Leaky ReLU slope	0.2				
Learning rate	0.0002				
Optimizer	Adam, $\beta_1 = 0.5, \beta_2 = 0.99$				
BatchNorm	$\epsilon = 0.00001$, $\beta = 0.98$				
bias initialization	0				

5.3.3 Evaluation

We followed the evaluation criteria introduced by the challenge organizers BraTS2018 [1]. Moreover, we evaluate impact of proposed mutual losses regarding handling imbalanced issues by precision-recall trade off.

The goal of segmentation is to delineate different tumor structures such as active tumorous core, enhanced tumorous, and whole tumorous. More than 200 teams attended to the BraTS 2018 challenge which Table 5.3 shows comparison results between top three groups obtained by the organizer.

From Table 5.3, the 3D DM-GAN achieved better results for whole tumor segmentation in terms of Dice, compared to the 2D DM-GAN. Except different discriminator loss, the network are different in CNN type. CNNs for segmentation can be categorized based on the dimension of convolutional kernel that is utilized. 2D CNNs use 2D convolutional kernels to predict the segmentation map for a single slice. Segmentation maps are predicted for a full volume by taking predictions one slice at a time. The 2D convolutional kernels are able to leverage context across the height and width of the slice to make predictions. However, because 2D CNNs take a single slice as input, they inherently fail to leverage context from adjacent slices. Voxel information from adjacent slices may be useful for the prediction of segmentation maps. 3D CNNs address this issue by using 3D convolutional kernels to make segmentation predictions for a volumetric patch of a scan. The ability to leverage interslice context can lead to improved performance but comes (see represented results on Figures 5.6 and 5.7).

Based on Table 5.3, we see the comparison and performance of our method in detail with other related approaches as well as winners in terms of Dice, Hausdorff distance, and sensitivity. Based on Table 5.3, Nvidia team [107] achieved best results for whole tumor segmentation with the ten cascade architecture including 3D U-Net, 3D-FCN, 3D-DenseNet, and applying variation intensity normalization in each network. Wang et al. [103] achieved the second rank of a challenge with a triple cascade of 3D U-Net architecture where each network is trained on three different regions of the tumor.

[1] http://www.med.upenn.edu/sbia/brats2018/evaluation.html

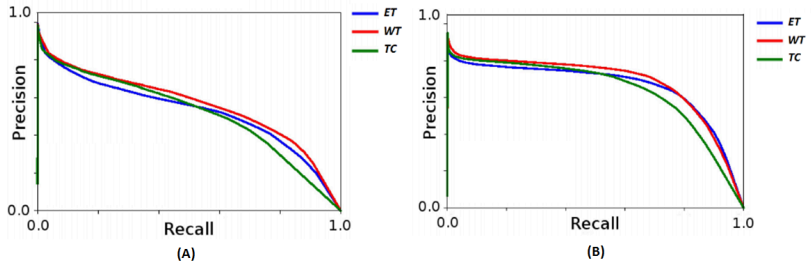


Figure 5.4: Comparison of Precision-Recall curves obtained by the cGAN and 2D recurrent deep mutual GAN approaches for semantic segmentation of three different brain tumor region. The PR curve examined by cGAN shown in the left side followed by PR curve achieved by 2D recurrent DM-GAN in the right side.

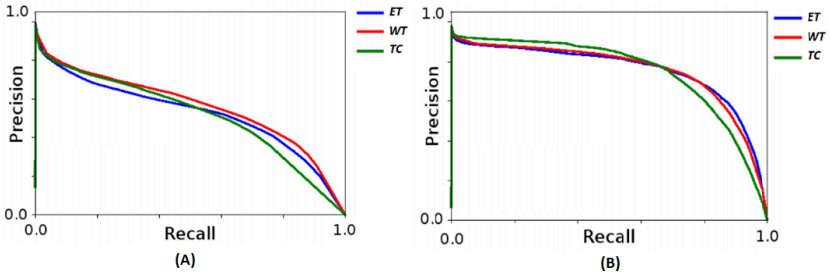


Figure 5.5: Comparison of Precision-Recall curves obtained by the cGAN and 3D deep mutual GAN approaches for semantic segmentation of three different brain tumor region. The PR curve examined by cGAN shown in the left side followed by PR curve achieved by 3D DM-GAN in the right side.

In order to evaluate the performance of semantic segmentation on handling imbalanced class problem, we use the Precision-Recall (PR) curve. Figures 5.4 and 5.5 summarize the trade-off between the true positive rate and the positive predictive value for a predictive model by cGAN and DM-GAN which the area under curve represent the success of our approach (see Figures 5.4-(B) and 5.5-(B)) for handling imbalanced class problem.

Table 5.3: Comparison results of our achieved accuracy for semantic segmentation by voxel-GAN (trained model with weighted loss) with related work and top ranked team, in terms of Dice and Hausdorff distance on five fold cross validation after 80 epochs while the reported results in second and third rows are after 200 epochs. WT, ET, and CT are abbreviation of whole tumor, enhanced tumor, and core of tumor regions respectively.

Model	Dice			Hausdorff		
	WT	ET	CT	WT	ET	CT
3D DM-GAN	0.91	0.80	0.85	3.25	4.1	6.24
2D DM-GAN	0.89	0.74	0.79	6.32	4.8	7.81
3D JointGAN [151]	0.87	0.68	0.84	5.62	6.8	9.81
RNN JointGAN [151]	0.86	0.68	0.82	5.87	7.1	10.04
cGAN	0.81	0.61	0.64	7.30	9.22	12.04
Ensemble of 10 3D-Models [102]	0.91	0.84	0.86	3.9	4.5	6.8
3D UNet + TTA [103]	0.88	0.79	0.78	4.5	5.9	8.0

Table 5.4: Comparison results of our achieved accuracy for semantic segmentation by voxel-GAN (trained model with weighted loss) with related work and top ranked team, in terms of sensitivity and specificity on five fold cross validation after 80 epochs while the reported results in second and third rows are after 200 epochs. WT, ET, and CT are abbreviation of whole tumor, enhanced tumor, and core of tumor regions respectively.

Model	Sensitivity			Specificity		
Methods	WT	ET	CT	WT	ET	CT
3D DM-GAN	0.90	0.89	0.87	0.99	0.99	0.99
2D DM-GAN	0.89	0.82	0.79	0.99	0.99	0.99
3D JointGAN [151]	0.88	0.75	0.78	0.99	0.99	0.99
RNN JointGAN [151]	0.86	0.74	0.77	0.99	0.99	0.99
cGAN	0.75	0.61	0.55	0.99	0.99	0.99

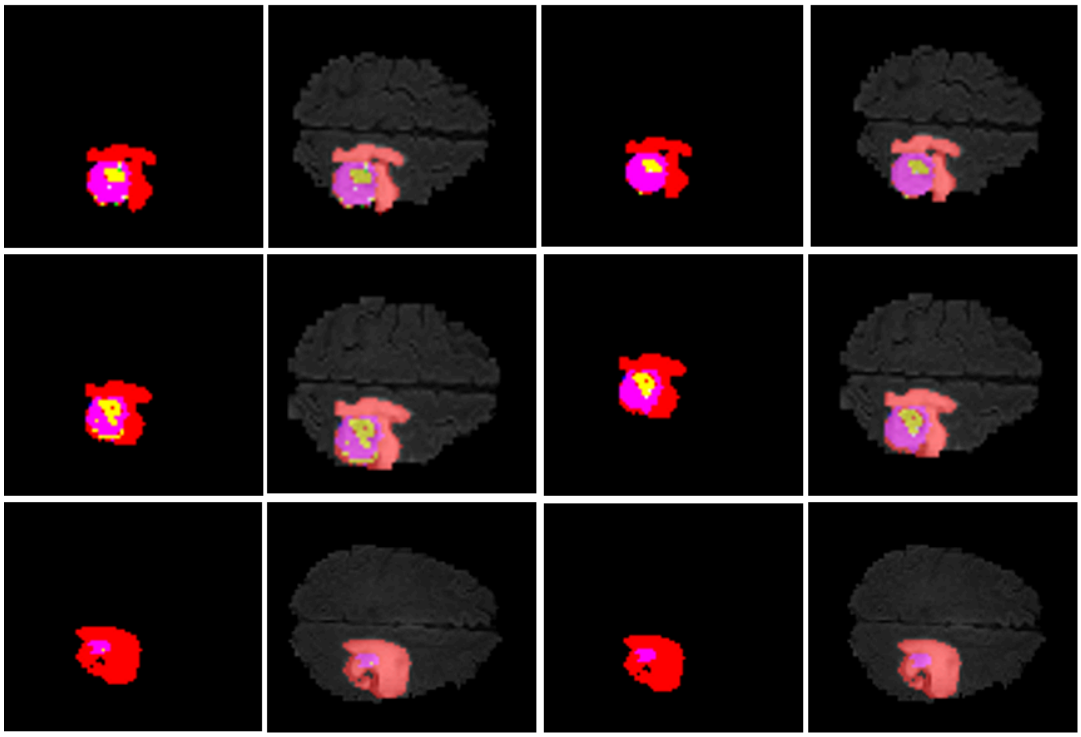


Figure 5.6: Predicted results from voxel GAN compared to deep mutual GAN model on axial views of and Brats18-CBICA-ALA.nz.120 from the test set overlaid T1C modality in second and forth column. First column shows the predicted results by deep mutual GAN while third show output by voxel-GAN. The red, pink, yellow color code the whole tumor (WT), the enhanced tumor (ET), and the tumorous core (TC) respectively.

5.4 Related Works

Recent studies shown [150, 153], mutually learned generator network achieve better results than generator trained by conventional distillation from a larger pre-trained discriminator. Furthermore, while the conventional understanding of distillation requires a more powerful discriminator than the intended generator, it turns out that in many cases, mutual learning of several large networks also improves performance compared to independent learning. This makes the deep mutual learning strategy generally applicable, e.g., it can also be used in application scenarios where there is no constraint on the model size, and the recognition accuracy is the only concern such as medical domain.

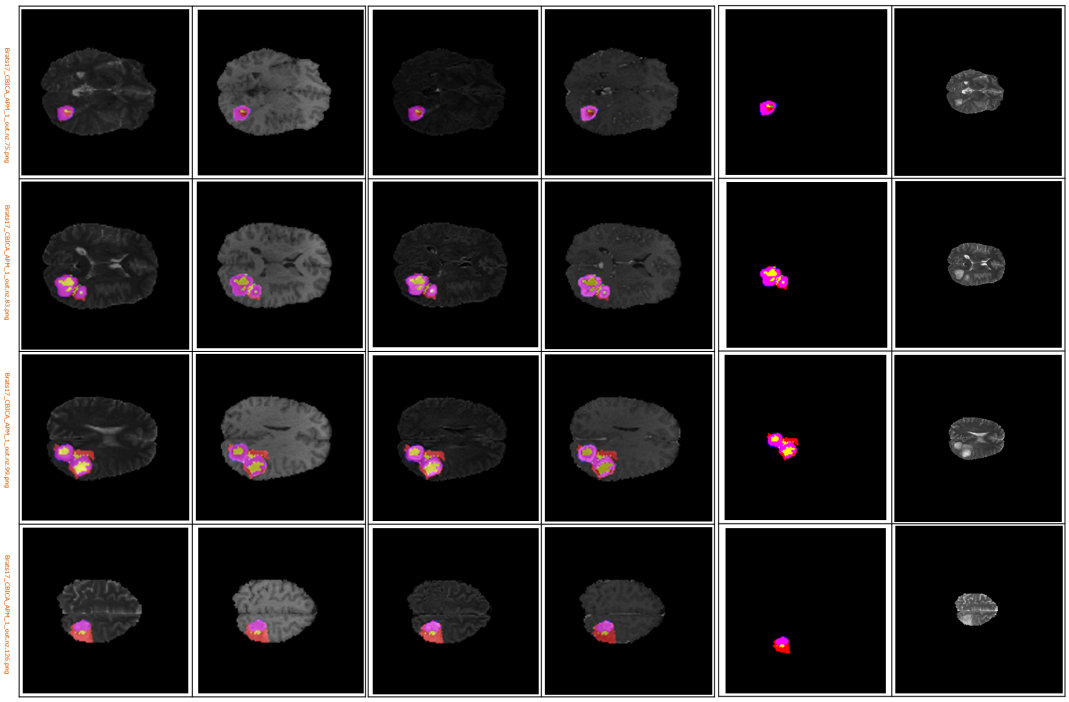


Figure 5.7: Predicted results from deep mutual GAN model. The red, pink, yellow color code the whole tumor (WT), the enhanced tumor (ET), and the tumorous core (TC) respectively.

As mentioned before, our proposed deep mutual GAN is built on top of DML method introduced by Zhang et al. [150] and our previous 3DJointGAN [151]. However, beside of different implementation and network architecture, in our framework, each discriminator substituted by different losses to mitigate the imbalanced class problem and minimize the generator error on a different aspect of quality. Unlike our previous 3DJointGAN [151], here the generator communicates and learns from each other through mimicry loss that aligns each generators class posterior with the class probabilities of other generator. Unlike 3DJointGAN, we train deep mutual GAN generator with same domain.

Other related ideas on collaborative learning include dual learning [154] where two cross lingual translation models teach each other interactively. But this only applies in this special translation problem where an unconditional within-language model is available to be used to evaluate the quality of the predictions. Furthermore, in dual learning different models have different learning tasks whilst

in mutual learning the tasks are identical.

Another related idea is model compression or distillation-based approach that has been proposed over a decade ago [155]. Knowledge distillation is a simple way to improve the performance of deep learning models on mobile devices. In this process, a large or complex network or ensemble model is trained to extract important features from the given data and can, therefore, produce better predictions. Then a small network is train with the help of the cumbersome model. This small network will be able to produce comparable results, and in some cases, it can even be made capable of replicating the results of the cumbersome network. Recently, Hinton et al. [156] explained the model distillation works due to the additional supervision and regularisation of the higher entropy soft-targets. Romero et al. [157] applied to distill powerful and easy-to-train large networks into small but harder-to-train networks. Likely, we transfer knowledge from a powerful network as discriminator to a small network, generator, in order to meet high performance and mitigate imbalanced class problem.

In contrast to the proposed model ensemble presented in the last chapter, we address dispensing the discriminator with a specific generator and allowing an ensemble of the generator to teach each other in mutual distillation.

5.5 Summary and Extensions

In this chapter, we addressed the problem of imbalanced data distribution by approaching impact of mutual information between independent ensemble model. We introduced deep mutual GAN (DM-GAN) composed by untrained generators and pre-trained discriminators. During the training process, each generator minimize predicted error by getting feedback from individual discriminator in adversarial setting. Additionally, each generator learns to collaborate with other generators during the training process which results shown the success for handling imbalanced data distribution. We presented the application of our proposed method on real patient data for semantic segmentation. In future, we plan to implement 2D and 3D deep mutual GAN architecture consisting more than four models in a distributed way.

5. LEARNING IMBALANCED SEMANTIC SEGMENTATION BY DEEP MUTUAL GANS

All proposed approaches in this thesis are not limited for application in the domain of medical diagnosis; they can be applicable where some classes have a significantly higher number of examples in the training set such as fraud detection, few-shot learning, and autonomous driving.

Bibliography

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 2018. xxvi, 65, 69, 75, 77, 78

[2] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10): 1993–2024, 2015. xxvi, 1, 29, 39, 65, 70, 103, 104, 126, 127

[3] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, JS. Kirby, JB. Freymann, K. Farahani, and Ch. Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nature Scientific Data*, 2017. xxvi, 39, 65, 103, 104, 126

[4] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and Ch. Davatzikos. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection. *The Cancer Imaging Archive*, 2017. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q. xxvi, 39, 65, 103, 104, 126

[5] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and Ch. Davatzikos. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collec-

tion. *The Cancer Imaging Archive*, 2017. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF. xxvi, 39, 65, 103, 104, 126

[6] Bruce Fischl, David H Salat, André JW Van Der Kouwe, Nikos Makris, Florent Ségonne, Brian T Quinn, and Anders M Dale. Sequence-independent segmentation of magnetic resonance images. *Neuroimage*, 23:S69–S84, 2004. 2

[7] Kilian M Pohl, John Fisher, W Eric L Grimson, Ron Kikinis, and William M Wells. A bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228–239, 2006. 2

[8] Stefan Bauer, Lutz-P. Nolte, and Mauricio Reyes. *Fully Automatic Segmentation of Brain Tumor Images Using Support Vector Machine Classification in Combination with Hierarchical Conditional Random Field Regularization*, pages 354–361. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23626-6. 2

[9] Antonio Criminisi, Jamie Shotton, Ender Konukoglu, et al. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2–3):81–227, 2012. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[11] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2, 25, 60, 84, 89, 92

[12] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 4

[13] Terry S Yoo, Michael J Ackerman, William E Lorensen, Will Schroeder, Vikram Chalana, Stephen Aylward, Dimitris Metaxas, and Ross Whitaker. Engineering and algorithm design for an image processing api: a technical report on itk-the insight toolkit. *Studies in health technology and informatics*, pages 586–592, 2002. 8

[14] Jiewei Jiang, Xiyang Liu, Kai Zhang, Erping Long, Liming Wang, Wangting Li, Lin Liu, Shuai Wang, Mingmin Zhu, Jiangtao Cui, et al. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. *Biomedical engineering online*, 16(1):132, 2017. 11

[15] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3):427–436, 2008. 11

[16] John O Awoyemi, Adebayo O Adetunmbi, and Samuel A Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9. IEEE, 2017. 11

[17] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *Advances in Neural Information Processing Systems*, pages 2255–2265, 2017. 11

[18] Zhengyuan Yang, Yixuan Zhang, Jerry Yu, Junjie Cai, and Jiebo Luo. End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perceptions. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2289–2294. IEEE, 2018. 11

[19] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, Mar 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0192-5. URL https://doi.org/10.1186/s40537-019-0192-5. 11

[20] N Japkowicz and S Stephen. The class imbalance problem: A systematic study intelligent data analysis, 2002. 11

[21] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 11

[22] Kandel Abraham and Bunke Horst. *Data Mining in Time Series and Streaming Databases*, volume 83. World Scientific, 2018. 11

[23] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*. Springer, 2018. 11

[24] Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4(6):962–969, 1993. 11

[25] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 4(9):1263–1284, 2008. 11

[26] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013. 11

[27] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015. 11

[28] Paula Branco andLuis Torgo and Rita Ribeiro. A survey of predictive modeling under imbal-anced distributions. *ACM Comput. Surv*, 49(2):1–31, 2016. 12

[29] Camelia Lemnaru and Rodica Potolea. Imbalanced classification problems: systematic study, issues and best practices. In *International Conference on Enterprise Information Systems*, pages 35–50. Springer, 2011. 13

[30] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. 13

[31] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017. 13, 53

[32] Simiao Yu, Hao Dong, Guang Yang, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, David Firmin, et al. Deep de-aliasing for fast compressive sensing mri. *arXiv preprint arXiv:1705.07137*, 2017. 13, 53

[33] Georgios Douzas and Fernando Bacao. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications*, 91:464–471, 2018. 21

[34] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018. 21

[35] J. Jang, T. Eo, M. Kim, N. Choi, D. Han, D. Kim, and D. Hwang. Medical image matching using variable randomized undersampling probability pattern in data acquisition. In *2014 International Conference on Electronics, Information and Communications (ICEIC)*, pages 1–2, Jan 2014. doi: 10.1109/ELINFOCOM.2014.6914453. 21

[36] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6): 1310–1321, 2018. 21

[37] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 21, 53

[38] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. In *Advances in Neural Information Processing Systems*, pages 1087–1098, 2018. 21, 22

[39] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 22

[40] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. 25, 69, 84, 98

[41] Yuan Xue, Tao Xu, Han Zhang, L. Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scalel1 loss for medical image segmentation. *CoRR*, abs/1706.01805, 2017. 25, 76, 84

[42] Pim Moeskops, Mitko Veta, Maxime W. Lafarge, Koen A. J. Eppenhof, and Josien P. W. Pluim. Adversarial training and dilated convolutions for brain MRI segmentation. *CoRR*, abs/1707.03195, 2017. URL http://arxiv.org/abs/1707.03195. 25, 84

[43] Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus H. Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *CoRR*, abs/1702.08014, 2017. 25, 84

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 26

[45] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL http://arxiv.org/abs/1502.03167. 26

[46] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005. 26

[47] Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1942–1950, 2017. 27, 53, 54

[48] Yeonkook J Kim, Yoonhwan Oh, Sunghoon Park, Sungzoon Cho, and Hayoung Park. Stratified sampling design based on data mining. *Healthcare informatics research*, 19(3):186–195, 2013. 27

[49] Mansour Keramat and Richard Kielbasa. A study of stratified sampling in variance reduction techniques for parametric yield estimation. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 45(5):575–583, 1998. 27

[50] Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014. 27

[51] Trong Duc Nguyen, Ming-Hung Shih, Divesh Srivastava, Srikanta Tirthapura, and Bojian Xu. Variance-optimal offline and streaming stratified random sampling. *arXiv preprint arXiv:1801.09039*, 2018. 27

[52] . URL http://www.medinfo.cs.ucy.ac.cy/index.php/downloads/datasets/. 29

[53] . URL http://www.isles-challenge.org/ISLES2016/. 29

[54] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017. 30

[55] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6629–6640, 2017. 30, 31

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 31

[57] Justin S. Paul, Andrew J. Plassard, Bennett A. Landman, and Daniel Fabbri. Deep learning for brain tumor classification. In *SPIE - Medical Imaging*, volume 10137, pages 1013710–1013710–16, 2017. 35

[58] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 34

[59] François Chollet et al. Keras, 2015. 40, 41, 66, 68, 97, 127, 129

[60] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org. 40

[61] Mina Rezaei, Haojin Yang, and Christoph Meinel. Recurrent generative adversarial network for learning imbalanced medical image semantic seg-

mentation. *Multimedia Tools and Applications*, pages 1–20, 2019. 40, 53, 59, 129

[62] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 41, 66, 68, 97, 127, 129

[63] Vasile Palade, Daniel-Ciprian Neagu, and Ron J Patton. Interpretation of trained neural networks by rule extraction. In *International Conference on Computational Intelligence*, pages 152–161. Springer, 2001. 42, 66

[64] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 42, 63, 67

[65] George E Nasr, EA Badr, and C Joun. Cross entropy error function in neural networks: Forecasting gasoline demand. In *FLAIRS Conference*, pages 381–384, 2002. 42, 60, 64, 67, 69, 98

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer International Publishing, 2015. 43, 48, 49, 60, 64, 86, 94, 111

[67] Klas EG Magnusson and Joakim Jaldén. A batch algorithm using iterative application of the viterbi algorithm to track cells and construct cell lineages. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 382–385. IEEE, 2012. 43, 111

[68] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Learning to detect cells using non-overlapping extremal regions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–356. Springer, 2012. 43, 111

[69] Saad Ullah Akram, Juho Kannala, Lauri Eklund, and Janne Heikkilä. Cell segmentation proposal network for microscopy image analysis. In Gustavo Carneiro, Diana Mateus, Loïc Peter, Andrew Bradley, João Manuel R. S. Tavares, Vasileios Belagiannis, João Paulo Papa, Jacinto C. Nascimento, Marco Loog, Zhi Lu, Jaime S. Cardoso, and Julien Cornebise, editors, *Deep Learning and Data Labeling for Medical Applications*, pages 21–29, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46976-8. 43, 111

[70] Patrick Ferdinand Christ, Florian Ettlinger, Felix Grun, Mohamed Ezzeldin A. Elshaer, Jana Lipkova, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Felix Hofmann, Melvin D'Anastasi, Seyed-Ahmad Ahmadi, Georgios Kaissis, Julian Holch, Wieland H. Sommer, Rickmer Braren, Volker Heinemann, and Bjoern H. Menze. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. *CoRR*, abs/1702.05970, 2017. URL http://arxiv.org/abs/1702.05970. 45, 46, 108

[71] Lei Bi, Jinman Kim, Ashnil Kumar, and Dagan Feng. Automatic liver lesion detection using cascaded deep residual networks. *CoRR*, abs/1704.02703, 2017. URL http://arxiv.org/abs/1704.02703. 45, 46, 107, 108

[72] Xiao Han. Automatic liver lesion segmentation using a deep convolutional neural network method. *arXiv preprint arXiv:1704.07239*, 2017. 45, 46

[73] Eugene Vorontsov, An Tang, Chris Pal, and Samuel Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *15th IEEE International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1332–1335, 2018. 45, 46, 107, 108

[74] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Isgum. Automatic segmentation and disease classification using cardiac cine mr images. *arXiv preprint arXiv:1708.01141*, 2017. 45, 48, 49, 77, 78

[75] Marc-Michel Rohé, Maxime Sermesant, and Xavier Pennec. Automatic multi-atlas segmentation of myocardium with svf-net. In *Statistical Atlases*

and Computational Modeling of the Heart (STACOM) workshop, 2017. 46, 48, 49, 77

[76] Clément Zotti, Zhiming Luo, Olivier Humbert, Alain Lalande, and Pierre-Marc Jodoin. Gridnet with automatic shape prior registration for automatic mri cardiac segmentation. *arXiv preprint arXiv:1705.08943*, 2017. 46, 48, 49, 77

[77] Fabian Isensee, Paul F Jaeger, Peter M Full, Ivo Wolf, Sandy Engelhardt, and Klaus H Maier-Hein. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 120–129. Springer, 2017. 48, 49, 77, 78

[78] Rudra PK Poudel, Pablo Lamata, and Giovanni Montana. Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 83–94. Springer, 2016. 48

[79] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *arXiv preprint arXiv:1809.06222*, 2018. 52

[80] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017. 53

[81] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 289–293. IEEE, 2018. 53

[82] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th International Sym-*

posium on Biomedical Imaging (ISBI 2018), pages 240–244. IEEE, 2018. 53

[83] Mina Rezaei, Haojin Yang, Konstantin Harmuth, and Christoph Meinel. Conditional generative adversarial refinement networks for unbalanced medical image semantic segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1836–1845. IEEE, 2019. 53

[84] Mina Rezaei, Haojin Yang, and Christoph Meinel. Whole heart and great vessel segmentation with context-aware of generative adversarial networks. In *Bildverarbeitung für die Medizin 2018*, pages 353–358. Springer, 2018. 53

[85] Yipeng Hu, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, Tom Vercauteren, J Alison Noble, and Dean C Barratt. Adversarial deformation regularization for training image registration neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 774–782. Springer, 2018. 53

[86] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–209. Springer, 2018. 53

[87] Mina Rezaei, Haojin Yang, and Christoph Meinel. Deep neural network with l2-norm unit for brain lesions detection. In *International Conference on Neural Information Processing*, pages 798–807. Springer, 2017. 53, 58

[88] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in Neural Information Processing Systems*, pages 5639–5649, 2017. 54, 55

[89] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *The European Conference on Computer Vision (ECCV)*, September 2018. 54

[90] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017. 58

[91] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *International MICCAI Brainlesion Workshop*, pages 64–76. Springer, 2017. 58

[92] Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P. Prabhu, Simon K. Warfield, and Ali Gholipour. Tversky as a loss function for highly unbalanced image segmentation using 3d fully convolutional deep networks. *CoRR*, abs/1803.11078, 2018. URL `http://arxiv.org/abs/1803.11078`. 58

[93] Mina Rezaei, Haojin Yang, and Christoph Meinel. voxel-gan: Adversarial framework for learning imbalanced brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 321–333. Springer, 2018. 58, 94

[94] Mina Rezaei, Haojin Yang, and Christoph Meinel. Generative adversarial framework for learning multiple clinical tasks. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018. 58

[95] Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, volume 3, page 435, 2003. 58

[96] Rudy Bonavia Inda, Maria-del-Mar and Joan Seoane. Glioblastoma multiforme: A look inside its heterogeneous nature. In *Cancer Archive 226-239*, 2014. 63, 107

[97] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005. 64, 69, 87, 97, 129

[98] Oskar Maier, Bjoern H Menze, Janina von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017. 65

[99] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, and Philippe Büchler. The virtual skeleton database: an open access repository for biomedical research and collaboration. *Journal of medical Internet research*, 15(11), 2013. 65

[100] Mina Rezaei, Konstantin Harmuth, Willi Gierke, Thomas Kellermeier, Martin Fischer, Haojin Yang, and Christoph Meinel. A conditional adversarial network for semantic segmentation of brain tumor. In *International MICCAI Brainlesion Workshop*, pages 241–252. Springer, 2017. 70, 71, 73

[101] Amir Gholami, Shashank Subramanian, Varun Shenoy, Naveen Himthani, Xiangyu Yue, Sicheng Zhao, Peter H. Jin, George Biros, and Kurt Keutzer. A novel domain adaptation framework for medical image segmentation. *CoRR*, abs/1810.05732, 2018. URL http://arxiv.org/abs/1810.05732. 70, 71

[102] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. *arXiv preprint arXiv:1810.11654*, 2018. 70, 132

[103] Guotai Wang, Wenqi Li, Sebastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. *arXiv preprint arXiv:1810.07884*, 2018. 70, 130, 132

[104] . URL braintumorsegmentation.org. 73

[105] Mahendra Khened, Varghese Alex, and Ganapathy Krishnamurthi. Densely connected convolutional network for cardiac mr image segmentation and heart diagnosis. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 140–151. Springer, 2017. 74, 78

[106] Irem Cetin, Gerard Sanroma, Steffen E Petersen, Sandy Napel, Oscar Camara, Miguel-Angel Gonzalez Ballester, and Karim Lekadir. A radiomics approach to computer-aided diagnosis with cardiac cine-mri. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 82–90. Springer, 2017. 74, 78

[107] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. *arXiv preprint arXiv:1711.01468*, 2017. 74, 76, 130

[108] Guotai Wang, Wenqi Li, Sebastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. *arXiv preprint arXiv:1709.00382*, 2017. 74, 76, 104, 105

[109] Pedro Amorim, chagas vinicius, and escudero guilherme. 3d u-nets for brain tumor segmentation in miccai 2017 brats challenge. In *Proceedings of the 6th MICCAI BraTS Challenge*, pages 9–14. BrainLes, 2017. 76

[110] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. *2017 International MICCAI BraTS Challenge*, 2017. 74, 76, 104, 105

[111] Yeonggul Jang, Yoonmi Hong, Seongmin Ha, Sekeun Kim, and Hyuk-Jae Chang. Automatic segmentation of lv and rv in cardiac mri. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 161–169. Springer, 2017. 77

[112] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2d and 3d deep learning techniques for cardiac

mr image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 111–119. Springer, 2017. 77

[113] Xin Yang, Cheng Bian, Lequan Yu, Dong Ni, and Pheng-Ann Heng. Class-balanced deep neural network for automatic ventricular structure segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 152–160. Springer, 2017. 77

[114] Robi Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012. 80

[115] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005. ISBN 0385721706. 80

[116] Gavin Brown. *Ensemble Learning*, pages 312–320. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_252. URL https://doi.org/10.1007/978-0-387-30164-8_252. 80

[117] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 80

[118] Chao Chen, Andy Liaw, Leo Breiman, et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110:1–12, 2004. 81

[119] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 81

[120] Harris Drucker. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115, 1997. 81

[121] Matteo Re and Giorgio Valentini. 1 ensemble methods: a review 3. 2012. 81

[122] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pages 90–95. IEEE, 1996. 81

[123] Liran Lerman, Romain Poussier, Olivier Markowitch, and François-Xavier Standaert. Template attacks versus machine learning revisited and the

curse of dimensionality in side-channel analysis: extended version. *Journal of Cryptographic Engineering*, 8(4):301–313, 2018. 81

[124] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. 81

[125] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003. 82

[126] Gareth M James. Variance and bias for general loss functions. *Machine Learning*, 51(2):115–135, 2003. 82

[127] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 311–320, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11726-9. 82

[128] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016. 83

[129] Andrew J Schaumberg, Mark A Rubin, and Thomas J Fuchs. H&e-stained whole slide image deep learning predicts spop mutation state in prostate cancer. *BioRxiv*, page 064279, 2018. 83

[130] Mehmet Günhan Ertosun and Daniel L Rubin. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1899. American Medical Informatics Association, 2015. 83

[131] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012. 83

[132] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 91

[133] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009. 100

[134] Spyridon(Spyros)Bakas. 2017 international miccai brats challenge. pages 1–352, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104

[135] Pawar et al. Residual encoder and convolutional decoder neural network for glioma segmentation. pages 219–225, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[136] Jin Zhua, Duo Wanga, Zhongzhao Tengb, and Pietro Lio. A multi-pathway 3d dilated convolutional neural network for brain tumor segmentation. pages 342–350, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[137] Alex et al. "brain tumor segmentation from multi modal mr images using fully convolutional neural network. pages 1–8, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[138] P. H. A Amorim, G. G Escudero, D. D. C Oliveira, S. M. Pereira, and A. A. Santos, H. M.and Scussel. 3d unets for brain tumor segmentation in miccai 2017 brats challenge. pages 9–14, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[139] Simon Andermatt, Simon Pezold, and Philippe Cattin. Multi-dimensional gated recurrent units for brain tumor segmentation. pages 15–19, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[140] Andrew Beers, Chang Ken, Brown James, Sartor Emmett, Mammen CP, Gerstner Elizabeth, Rosen Bruce, and Kalpathy-Cramer Jayashree. Sequential 3d unets for brain tumor segmentation. pages 20–28, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[141] Marcel Cata, Adria Casamitjana, Irina Sanchez, Marc Combalia, and Veronica Vilaplana. Masked v-net: an approach to brain tumor segmentation. pages 42–50, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104, 105

[142] Dong Shidu. A separate 3dsegnet architecture for brain tumor segmentation. pages 54–60, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104

[143] Zach Eaton-Rosen, Wenqi Li, Guotai Wang, Tom Vercauteren, Bisdas Sotirios, Sebastien Ourselin, and M. Jorge Cardoso. Using niftynet to ensemble convolutional neural nets for the brats challenge. pages 61–67, 2017. URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf. 104

[144] RG Rodríguez Colmeiro, CA Verrastro, and T Grosges. Multimodal brain tumor segmentation using 3d convolutional networks. In *International MICCAI Brainlesion Workshop*, pages 226–240. Springer, 2017. 104

[145] N Pawlowski, M Rajchl, M Lee, B Kainz, D Rueckert, and B Glocker. Ensembles of multiple models and architectures for robust brain tumour segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers*, volume 10670, page 450. Springer, 2018. 104, 105

[146] Xiao Han. Automatic liver lesion segmentation using A deep convolutional neural network method. *CoRR*, abs/1704.07239, 2017. URL http://arxiv.org/abs/1704.07239. 107, 108

[147] Rahil Shahzad, Shan Gao, Qian Tao, Oleh Dzyubachyk, and Rob vander Geest. *Automated Cardiovascular Segmentation in Patients with Congenital Heart Disease from 3D CMR Scans: Combining Multi-atlases and Level-Sets*, pages 147–155. Springer International Publishing, 2017. 113

[148] Lequan Yu, Xin Yang, Jing Qin, and Pheng-Ann Heng. *3D FractalNet: Dense Volumetric Segmentation for Cardiovascular MRI Volumes*, pages 103–110. 2017. 113

[149] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Dilated convolutional neural networks for cardiovascular mr segmentation in congenital heart disease. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 95–102. Springer, 2016. 113

[150] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018. 121, 133, 134

[151] Mina Rezaei, Haojin Yang, and Christoph Meinel. Learning imbalanced semantic segmentation through cross-domain relations of multi-agent generative adversarial networks. In *Medical Imaging 2019: Computer-Aided*

Diagnosis, volume 10950, page 1095027. International Society for Optics and Photonics, 2019. 121, 132, 134

[152] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 122, 123, 124, 126

[153] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR 2019*, 2019. 133

[154] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016. 134

[155] Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM, 2006. 135

[156] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 135

[157] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 135

Acronyms

ACDC Automated Cardiac Diagnosis Challenge

BraTS Brain Tumor Segmentation

cGAN Conditional Generative Adversarial Network

CMRI Cardiovascular Magnetic Resonance Images

CNN Convolutional Neural Network

CT Computed Tomography

DCGAN Deep Convolutional Generative Adversarial Network

DL Deep Learning

ED End of Diastolic

ES End of Systolic

ET Enhanced Tumor

FID Fréchet Inception Distance

FLAIR

GAN Generative Adversarial Network

HGG High Grade Glioma

HVSMR whole-Heart and great Vessel Segmentation from 3D cardiovascular MRI in Congenital Heart Disease

ILSVRC Imagenet Large Scale Visual Recognition Challenge

IPMI international conference on Information Processing in Medical Imaging

ISBI IEEE International Symposium on Biomedical Imaging

ISLES Ischemic Stroke Lesion Segmentation

KL KullbackLeibler

lad least absolute deviations

LGG Low Grade Glioma

LiTS Liver Tumor Segmentation

LV Left Ventricle

MI Mutual Information

MICCAI international Conference on Medical Image Computing and Computer Assisted Intervention

MIDL International Conference on Medical Imaging with Deep Learning

MRI Magnetic Resonance Imaging

MS Multiple Sclerosis

MYO Myocardium

ReLu Rectified Linear Unit

RNN-GAN Recurrent Generative Adversarial Network

ROI Region of Interest

RV Right Ventricle

SMOTE Synthetic Minority Over-sampling Technique

SPIE The International Society for Optics and Photonics

TC Tumor Core

US Ultra Sound

VAE Variational Auto Encoder

VOE Volume Overlap Error

WT Whole Tumor

Fluid Attenuated Inversion Recovery