Fakultät für Informatik
Technische Universität München

Deep Learning for Medical Image Analysis

Xiaobin Hu

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Christian Mendl

Prüfende der Dissertation:
1. Prof. Dr. Björn H. Menze
2. Prof. Dr. Vasileios Belagiannis

Die Dissertation wurde am 26.04.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 18.10.2021 angenommen.

Abstract

As one of the most powerful tools in the machine learning community, deep learning methods have demonstrated record-breaking success in previously challenging tasks such as computer vision, natural language processing, speech recognition, and social multimedia retrieval.

Different from other traditional machine learning algorithms relaying on a hand-crafted way by well-experienced researchers, deep learning is capable of mimicking the workings of the human brain in automatically processing data and extracting features for use in decision making. Recently, considering the enormous amounts of data containing valuable knowledge for clinical diagnosis and treatment are generated, medical experts have become fatigued to make interpretations and have begun to benefit from deep learning-based computer-assisted interventions. Essentially, deep learning allows the radiologists for less labor-intensive intervention and mines the informative representations from sparse and noisy data in a self-taught approach only requiring little effort on data preprocessing. The applications make it possible for non-experts to utilize deep learning for specific clinical-relevant studies, especially in medical image analysis. Compared with the unprecedented success of deep learning in general image analysis, the improvement of medical image analysis is still in the early stage due to the complexity of medical imaging. There are some domains, including the diversity of network architectures, training strategy, network interpretability, etc., which should be further explored and designed according to specific clinic tasks.

This thesis aims to develop effective and novel deep learning based algorithms to resolve lesion segmentation, disease prognostic analysis issues, and medical image synthesis, such as brain glioma multi-class segmentation, natural killer/T cell lymphoma multi-stage segmentation, prognostic analysis of natural killer/T cell lymphoma, and MR image enhancement. More specifically, (1). To incorporate the nesting topological priors among whole tumor, tumor core, and active tumor, we present a multi-level activation function embedded in 3D residual U-Net architecture for hierarchical multi-class segmentation. (2). Considering that Extranodal natural killer/T cell lymphoma (ENKL) segmentation is crucial for clinical decision support

and treatment planning, we propose an automatic and coarse-to-fine approach for ENKL segmentation using adversarial networks. (3). Due to a low-survival rate and difficult prognostic prediction of ENKL disease, we develop a weakly supervised deep learning (WSDL) method that could utilize incomplete/missing survival data to predict the prognosis of extranodal natural killer/T cell lymphoma. We build a positive–negative unlabeled (PNU) classifier to generate implicit labels for incomplete survival data and then retrain deep convolutional neural networks with labeled and unlabeled data to obtain the final prognosis. 4). To remove common distortions (e.g., artifacts, blur, and noise) of degraded MRI via learning the symmetry and self-similarity relationship of patch-based features in multi-modal brain MR images where the structure of the brain is normally symmetry, we designed a specialized Graph-based structure to merge the high-similarity information of sub-regions by updating larger weights to the more important and similar nodes or features in a graph attention fashion.

Zusammenfassung

Als eine der leistungsstärksten Methoden des maschinellen Lernens hat Deep Learning hervorragende Resultate in anspruchsvollen Gebieten wie Computer Vision, Natural Language Processing, Spracherkennung und Social Multimedia Retrieval erzielt.

Anders als konventionelle Ansätze des maschinellen Lernens, die auf manuelle Weise von erfahrenen Wissenschaftlern erstellt wurden, ist Deep Learning dazu fähig, die Arbeitsweise des menschlichen Gehirns nachzuahmen und so automatisch Daten zu verarbeiten und charakteristische Merkmale zu extrahieren. In Anbetracht der enormen Menge an wertvollen Daten, die bei einer klinischen Diagnose und Behandlung erstellt werden, haben medizinische Experten damit begonnen, die Vorteile von Deep Learning angewandt auf medizinische Problemstellungen zu nutzen. Dies ermöglicht Radiologen weniger arbeitsintensive Untersuchungen und benötigt nur wenig Aufwand in der Vorbereitung der Daten, um informative Repräsentationen aus lückenhaften und ungenauen Daten eigenständig zu erlernen. Die Anwendungen ermöglichen es auch Nicht-Experten, Deep Learning für spezifische klinisch relevante Studien, besonders im Bereich der medizinischen Bildverarbeitung, zu nutzen. Verglichen mit dem Erfolg von Deep Learning in der allgemeinen Bildverarbeitung sind die Fortschritte auf dem Gebiet der medizinischen Bildverarbeitung noch immer in einem Frühstadium aufgrund der Komplexität medizinischer Bildgebung. Teilbereiche wie beispielsweise die Architektur des Netzwerks, die Trainingsstrategie oder die Interpretierbarkeit des Netzwerks sollten dabei weiter untersucht und entsprechend für klinische Anwendungen angepasst werden.

Diese Arbeit beschäftigt sich damit, effektive neue Algorithmen basierend auf Deep Learning zur Anwendung in Läsionssegmentierung, prognostischer Krankheitsanalyse und medizinischer Bildsynthese zu entwickeln. Insbesondere handelt es sich dabei um Multiklassen-Segmentierung von Gehirn Gliomen, mehrstufige Segmentierung von Natürlichen Killer-T-Zell-Lymphomen, prognostische Analyse Natürlicher Killer-T-Zell-Lymphome und MRT Bildverbesserung. Spezifischer beschäftigt sich diese Arbeit mit (1). Um jeweils den topologischen Prior des ganzen Tumors, des Tumorkerns und des aktiven Tumors miteinzubinden, stellen wir eine mehrstufige

Aktivierungsfunktion integriert in eine 3D Residual U-Net Architektur für hierarchische Multiklassen-Segmentierung. (2). In Anbetracht dessen, wie entscheidend die Segmentierung der extranodalen Natürliche Killer-T-Zell-Lymphome (ENKL) für die klinische Entscheidungsunterstützung und Behandlungsplanung ist, stellen wir einen automatischen coarse-to-fine Ansatz für ENKL Segmentierung mittels adversarialen Netzwerken vor. (3). Aufgrund der niedrigen Überlebensrate und schwieriger Prognose der ENKL Erkrankung haben wir eine Weakly Supervised Deep Learning (WSDL) Methode entwickelt, die unvollständige oder fehlende Überlebensdaten verwenden könnte, um Prognosen für ENKL zu erstellen. Wir verwenden einen Positiv-Negativ Unlabeled (PNU) Klassifikator, um implizite Labels für unvollständige Überlebensdaten zu generieren. Mit diesen trainieren wir erneut Deep Convolutional Neural Networks mit vollständigen und unvollständigen Daten, um die finale Prognose zu erhalten. (4). Um häufige Störungen wie Artefakte, Unschärfe und Rauschen aus fehlerhaften MRT-Scans zu entfernen, nutzen wir die symmetrische und selbstsymmetrische Beziehung von Regions-Merkmalen aus MRT-Scans des Gehirns, in dem normalerweise eine symmetrische Struktur vorliegt. Dafür haben wir eine spezielle graph-basierte Struktur entworfen, die, im Stile von Graph Attention, Informationen von Regionen mit hoher Ähnlichkeit zusammenführt, mit höherer Gewichtung für wichtigere und ähnlichere Knoten und Merkmale.

Acknowledgements

First and foremost I am extremely grateful to my supervisor, Prof. Bjoern H. Menze, who gave me an opportunity for Ph.D. study in the field of medical image analysis and deep learning. His immense knowledge and selfless supports have encouraged me to overcome some barriers in my academic research. It is my great honor to work with him in a flexible environment for around four years research journey. I greatly appreciate that he gave me lots of invaluable advice, continuous support, and patience during my research.

Additionally, I would like to thank Dr. Kuangyu Shi and Dr. Marie Piraud who supervised me for some projects. I would also like to express my gratitude to Prof. Wenqi Ren and Qi Dou for their research guidance in our collaborative projects. I would also like to thank each co-author of my publications, who give me a lot of valuable suggestions and feedback including technical writing and algorithms. Hongwei Li helped me a lot in the early stage of my Ph.D. study. Yu Zhao generous shared me with research experience which largely broadened my horizon. Diana Waldmannstetter patiently polished my paper in the aspect of languages and organization. John LaMaster and Carolin Pirkl made a great contribution to recording an online conference video during the COVID-19 period. Anjany Sekuboyina gave me some meaningful ideas that inspired me a lot. Lina Xu patiently taught me how to operate medical software and tool for medical image registration and Dolp Andreas helped me in writing my dissertation abstract.

Furthermore, I also expressed my gratitude to our colleagues and friends at the image-based biomedical modeling group, Amir Hossein Bayat, Ivan Ezhov, Giles Tetteh, Suprosanna Shit, Johannes Paetzold, Fernando Navarro, Oliver Schoppe, Judith Zimmermann, Xin Liu, Dhritiman Das, and Timo Löhr. Every communication with me left me unforgettable memories in Germany.

I would like to acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for my research. I also give our gratitude to the financial support of China Scholarship Council (CSC).

Last but not least, my parents (Jinlu Hu and Lanfang Wu) and sister Xia Hu always stand there to give me unconditional support and love. Thanks my girlfriend Ruolin Shen for accompanying me through my study.

Contents

List of Figures

Acronyms

AI . . . . . artificial intelligence

CNN . . . . . convolutional neural network

CPUs . . . . . central processing units

CRFs . . . . . conditional Random Fields

CT . . . . . computed tomography

DCNNs . . . . . deep convolutional neural networks

ENKTL . . . . . Extranodal natural killer/T cell lymphoma, nasal type

FCN . . . . . fully convolutional network

GANs . . . . . generative Adversarial Networks

GNNs . . . . . graph Neural Networks

GPUs . . . . . graphics processing units

LR . . . . . low-resolution

MC . . . . . Monte Carlo

MLP . . . . . multi-layer perceptron

MRI . . . . . magnetic resonance imaging

NAS . . . . . neural architecture search

PET . . . . . positron emission tomography

PFS . . . . . progression-free surviva

PNU . . . . . positive–negative unlabeled

PSI . . . . . prediction similarity index

SR . . . . . super-resolution

TLG . . . . . total lesion glycolysis

WSDL . . . . . weakly supervised deep learning

1

Introduction

Recently, a great explosion of medical image data attracts significant attention from the data-driven research community but also brought massive burdens for radiologists to make clinical-relevant interpretations. These medical imaging techniques such as computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and X-ray are widely used into the stage of detection, diagnosis, treatment and prognostic analysis [1]. Considering the labor-intensive character of radiologists and the huge cost to train medical experts, it is a great challenge to analyze all collected data within limited time in a human intervention way. There exist urgent demands and strong desires for computer-assisted intervention technique which can alleviate the workload of medical experts and give an efficient and personalized diagnosis or treatment. Although the development progress in computational medical image analysis is not satisfactory, the appearance of machine learning techniques has greatly changed this situation. It mainly attributes to the fact that machine learning can learn and capture the informative features that represent the regularities or patterns of data and play a key role in different tasks of medical image analysis. However, the features extraction was implemented by medical experts on the basis of task-relative knowledge which was a difficult barrier for non-experts to explore machine learning algorithms for their own researches [2].

As a most recent technique of machine learning, deep learning [3] improves the feature extraction stage to an advanced learning stage from data. Specifically, it can automatically extract and discover informative features from source data and largely minimize human intervention only requiring minor preprocessing [4], which is different from the previous features extraction in a handcrafted manner based on domain-specific knowledge. The self-learning manner of deep learning requires massive computational resources which are available nowadays due to the fast development of computing power (i.e., (high-tech central processing units (CPUs) and graphics processing units (GPUs)). The generation of a huge amount of medical data also boosts the performance of deep learning to an unprecedented level on different tasks of

medical image analysis [5, 6, 7, 8, 9]. Recently, reference [10] presents a deep learning system based on a large representative dataset from the UK and USA and shows this system is capable of outperforming human experts and decreasing absolute errors of 5.7% and 1.2% in false positives and 9.4% and 2.7% in false negatives in breast cancer diagnosis. Besides, a study [11] reported that the proposed deep learning algorithm can achieve comparable or superior performance than radiologists using 6,716 National Lung Cancer Screening Trial cases to predict the risk of lung cancer.

Since the data type of medical image accounts for more than 90% medical data [12], the medical image analysis plays an irreplaceable role in relevant clinical stages, such as diagnosis, treating plan, and prognostic prediction. Although recent researches have achieved some progress in these aspects, there are still some major challenges and bottlenecks that we should overcome. Firstly, considering the various geometry or clinical characters of different diseases, a tailored deep learning network embedding these specific disease priors should be considered and constructed in priority. Secondly, for label-scarcity diseases, an efficient semi-supervised or unsupervised system should be developed to well exploit unlabeled data for performance improvement. Thirdly, considering the lack of sufficient interpretability for decision making in deep learning models, the networks investigating the interpretability and uncertainty in clinical practice should be further analyzed to boost the widespread acceptance of model's decision. In our thesis, we aim to leverage state-of-the-art techniques to resolve all the above bottlenecks and challenges, especially in segmentation tasks and prognostic analysis.

1.1 Medical Image Segmentation

Since segmentation results provide the vital priors of the disease, such as lesion location and the lesion volume size, we regard the segmentation as the foundation of quantitative image analysis. However, manual segmentation is a tedious, time-consuming, and labor-intensive task and requires knowledge from bio-medical experts. There exists a great need for a fast and robust automated segmentation algorithm. The technology can benefit plenty of downstream tasks (e.g., lesion quantification, disease diagnosis, treatment planning, surgery monitoring, and navigation [13, 14, 15, 16, 17, 18, 19]). Specifically, after lesion quantification by segmentation, the diseases are evaluated and cataloged into different stages based on the lesion-quantification knowledge and other diagnoses information. For each stage, there exists a corresponding suitable treatment planning. Furthermore, the data fusion based on the lesion priors from segmentation and others (e.g. patients' age, surgery performance, recovery after surgery) is used to predicate the survival time and the risk of the disease recurrence. Although there are some successful applications in medical image segmentation, it is still a challenging task after considering the following causes:

- Low contrast characteristic: Low contrast medical images are more difficult to segment the boundaries of lesion or anatomical structural in comparison with high-contrast natural images.

- Noise: Random noise generated in the scanning process by motion disturbs the uniformity of pixel-based intensity.

- Imbalance problem: Imbalance samples between foreground and background lead to the ignorance of small lesions.

- Lack of label data: Label scarcity is a tough challenge for data-driven algorithms to learn segmentation knowledge due to the costly and time-consuming annotations by radiologists.

- Shape variability: Shape variability of different lesions or organs varies largely, which impeding the proposal of unified neural architecture for multiply segmentation tasks.

- Annotation Uncertainty: The annotations implemented by different radiologists are not identical and this uncertainty characteristic weakens the interpretability of neural network architecture and deteriorate the overall performance.

Recently, some studies including rule-based [20, 13, 21, 22], atlas based [23], machine-learning based [24, 25, 26, 27] and deep-learning based segmentation algorithms [28, 29] have proposed to tackle above problems. As a dominant area of deep-learning based methods, convolutional neural network (CNN) and their variants [28, 30, 31, 32] have been verified to be extremely effective for informative feature extractions used in a variety of semantic segmentation tasks [29]. Another universal architecture are encoder-decoder framework mainly consisting of down-sampling path, up-sampling path, and skip connections, such as 2D U-net [33], 3D U-net [34, 35], and their variants [36, 37, 38, 39]. This network consists of a contracting path to capture context and a symmetric expanding path that enables precise localization for segmentation. Although all these CNN methods have achieved promising results, they suffer from the limitation of insufficiently learning both local and global contextual information between pixels. Therefore, models such as the conditional Random Fields (CRFs) are implemented to embed the spatial contiguity in the output maps [40]. Some generative Adversarial Networks (GANs) [41, 42] incorporating a multi-scale L1 loss function are proposed to force the network to learn both global and local features, for capturing long- and short-range spatial relationships between pixels.

1.2 Medical Image Diagnosis

Medical imaging is crucial in diagnosing the various types of diseases among patients across the healthcare system [43]. The medical images including MRI, CT, Ultrasound, and X-Rays are used for disease diagnosis implemented by radiologists in a manual manner. But diagnostic errors occur when radiologists are not well-trained or limited by their time or attentions caused by numerous patients. Thus, some researchers aims to explore computer-aided diagnosis to assist the radiologists in disease diagnosis [44, 45, 46]. Considering the limitation of machine learning in features extraction which requiring well-skilled or well-experienced knowledge for clinicians, a deep learning diagnosis system attracts a lot of attention due to its capability in automatic feature selection. Specifically, deep neural networks can automatically learn informative knowledge without human interventions from medical image data. Deep learning has proven to be advantageous for computer-aided diagnosis in medical imaging, such as for the differential diagnosis of coronavirus disease 2019 [47], skin cancer [48], and diabetic retinopathy [49]. Moreover, it has been developed to help identify imaging-based biomarkers, leading to an improvement in the prognosis of, for example, lung cancer [50, 51], gliomas [52], and nasopharynx cancer [53].

Although the development of deep learning in diagnosis or prognostic analysis depends on the availability of a huge amount of data, it is usually challenging to

gather a large cohort of patients with survival follow-up after administering the same therapeutic regime. Clinical trials are often associated with incomplete or missing follow-up due to factors such as insufficient follow-up time, patient tolerance, and compliance. This consequently hampers the extensive development of deep learning methods for predicting therapeutic prognosis. Maximizing the utility of data gathered by clinical trials is thus a key area of research, which should be explored and resolved to further improve the diagnosis or prognosis results.

1.3 Medical Image Synthesis

Medical images have been widely used in clinics to provide visual representations for disease diagnosis, treating plan, and prognostic analysis. But it is an inevitable dilemma to achieve a balance between image resolution, signal-to-noise ratio, and acquisition time [54]. For magnetic resonance imaging (MRI) sequences, higher resolution imaging grasps more structural details and provides more diagnostic information, but requires longer acquisition time [55]. Since the signal-to-noise ratio is proportional to the slice thickness and the square root of scanning time, the longer acquisition time leads to the performance drop of the signal-to-noise ratio and tends to generate artifacts caused by physiologic motion such as respiratory motion and physical movement of subjects. Considering the limited and costly MRI resource, some thick slices and low scan time MRI images are usually utilized to get a desired signal-to-noise ratio [56, 57, 58]. Consequently, the use of image synthesis techniques to enhance medical image quality is an established field of research in medical image computing and imaging physics [59], for example, to prevent blurring and information loss when co-aligning different image volumes in a multi-parametric sequence. Besides, due to modality corruption, incorrect machine settings, allergies to specific contrast agents, and limited available time, it is often not guaranteed to obtain complete set of MRI sequences to provide rich information for clinical diagnosis and therapy. In this regard, development of cross-modality or cross-protocol MRI synthesis techniques is important to homogenize and "repair" such real-world data collections via efficient data infilling and re-synthesis, and make them accessible for algorithms that require complete data sets as input.

Recently, convolutional neural network (CNN) based approaches have shown dramatic improvements and exhibited state-of-the-art performance in image synthesis. For image enhancement, some studies [60, 61, 62] uses CNN to learn the mapping between low-resolution and high-resolution images. But considering the recovering more satisfactory level of image realism, generative Adversarial Networks (GANs) [63] and its variants [64, 65, 66, 67, 68, 69, 70, 71, 72] are proposed to rehabilitate more

faithful images. GANs consist a generator network to generate synthesized images and a discriminator network to discriminate the real or synthesized images. The generator and discriminator are alternately trained by back propagation in an adversarial fashion in a min-max game. GANs are extensively used in 7T MRI synthesis from 3T MRI images [73], PET-CT translation [74], and MRI cross-modalities re-synthesis [75, 76]. But some networks fail to exploit global structural information and self-similarity details of medical images and are effective for only specific tasks.

1.4 Summary of Contributions

This thesis aims to explore some challenging issues and advanced networks in the tasks of lesion segmentation, disease diagnosis, and medical image synthesis. Specifically, we focus on the two segmentation subtasks (i.e., the multi-class segmentation of glioma and the segmentation of Extranodal natural killer/T cell lymphoma, nasal type (ENKTL), prognostic diagnosis of ENKTL using the incomplete follow-up data, and MRI enhancement based on graph neural network using self-similarity features between nodes.

Chapter 3: Hierarchical Multi-class Segmentation of Glioma Images Using Networks with Multi-level Activation Function

For many segmentation tasks, especially for the biomedical image, the topological prior is crucial information that is useful to exploit. As the most common family of brain tumors, Glioma forms some of the highest mortality and economically costly diseases of brain cancer [77, 78, 79]. The containment/nesting is a typical inter-class geometric relationship in MICCAI brain tumor (glioma) segmentation challenge with its three hierarchically nested classes 'whole tumor', 'tumor core', 'active tumor'. While CNN segmentation algorithms are abundant in biomedical imaging, only very few make use of nested-topological prior information. Among the few that do [80, 81, 16, 82, 83, 84, 85], we find three different approaches. First, the use of cascaded algorithms where the network consists of successive segmentation networks. Second, the information on the nested-classes is incorporated into the loss function, imposing penalties on solutions that do not respect the nested geometry relations. Third, Markov random fields are used to formalizing class relationships in the post-processing of the network output.

Here, we make use of a new activation function [86] that is directly implementing class hierarchy in the network training and generalize it to 3 nested classes. For the

glioma labels, we assume that active tumor regions are always contained in the tumor core which is surrounded by the tumor edema, resulting in a hierarchical three-class model. In sharp contrast with nested-class method, the softmax-based method of multi-class ignores the geometric prior between different classes, and assumes the classes are mutually exclusive, meaning one pixel cannot belong to different classes at the same time, which absolutely discards the topological information and sometimes leads the unreasonable segmentation results. The nested classes relationship is introduced into the 3D-residual-Unet architecture [87]. The network comprises a context aggregation pathway and a localization pathway, which encodes increasingly abstract representation of the input as going deeper into the network, and then recombines these representations with shallower features to precisely localize the interest domain via a localization path.

The model is trained on the training dataset of Brats2018 [14, 88, 89], and 20% of the dataset is regarded as the validation dataset to determine parameters. When the parameters are fixed, we retrain the model on the whole training dataset. The performance achieved on the validation leaderboard is 86%, 77% and 72% Dice scores for the whole tumor, enhancing tumor and tumor core classes without relying on ensembles or complicated post-processing steps. Based on the same start-of-the-art network architecture, the accuracy of nested-class (enhancing tumor) is reasonably improved from 69% to 72% compared with the traditional Softmax-based method which blind to topological prior. The comparison of Dice score criteria indicates the nested-class method achieves higher accuracy than the softmax-based method, especially for the internal-classes.

Chapter 4: Coarse-to-Fine Adversarial Networks and Zone-Based Uncertainty Analysis for NK/T-Cell Lymphoma Segmentation in CT/PET Images

Extranodal natural killer/T cell lymphoma, nasal type (ENKTL) is a kind of rare disease with a low survival rate that primarily affects Asian and South American populations. ENKTL occurs predominantly in the nasal, paranasal and oropharyngeal sites. 18F-FDG PET/CT scanning is currently the most effective imaging modality for staging, monitoring response, and predicting prognosis for many kinds of lymphomas [90]. Several investigations identified that almost all ENKTL are FDG avid [91, 92]. The segmentation difficulties in the ENKTL dataset mainly stem from three aspects: 1) The large variations in the shape, size and location of the lymphoma. 2) Due to the large image sizes, network suffers from memory size, complicating image processing approaches that take the whole volume as input. 3) The images coming

from PET and CT are not identical in their sizes. Consequently, this complicates the straightforward usage of information from two modalities for boosting segmentation accuracy

For pixel-wise semantic segmentation, CNNs have also achieved remarkable successes. Different models, such as fully convolutional network (FCN) [28], encoder-decoder structure [34], conditional random fields [40], U-Net [33], cascade architectures [93], and 3D CNN [94], were proposed to segment pixel-level or voxel-level instances. Recently, Xue et al. [41] employed a GAN-based network which combined with multi-scale loss function to learn global and local features. However, it requires more computational cost and memory for the network, when the image size gets too large. Additionally, there is the problem of label and class imbalance, which deteriorates the segmentation results. To mitigate these problems, we propose the coarse-to-fine adversarial network for ENKTL segmentation, which achieves high segmentation accuracy by locating lesion zones, while cropping unnecessary information reduces computational cost.

To the best of our knowledge, this paper is one of the first deep learning studies on computer-aided diagnosis systems for an ENKTL dataset. The coarse stage acts as a dimensionality reduction to roughly locate the lesion bounding box and crops redundant information to facilitate the fine segmentation, which is crucial to reduce segmentation time and avoid memory problems. The fine segmentation is an end-to-end adversarial network with a generator and a discriminator part. Spatial context information and hierarchical features are exploited by introducing a multi-scale L1 loss function in both the generator and discriminator parts without further smoothing of predicted label maps using CRFs. Further, we present an exploration of zone-based uncertainty estimates based on Monte Carlo (MC) dropout [95, 96, 97] in the context of deep networks for medical image segmentation. This uncertainty analysis can give a clear understanding of the main source of uncertainty in the respective zones and is crucial for a quantitative evaluation of an algorithm's stability. It also makes it possible to permit subsequent optimization by engineers and revision by clinicians.

Chapter 5: Weakly supervised deep learning for determining the prognostic value of 18F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type

Extranodal natural killer/T cell lymphoma, nasal type (ENKTL) is a rare type of lymphoma with poor survival outcome [98, 99, 98]. It constitutes $\leq$1% of all lymphomas in Western countries and 3–9% of all malignant lymphomas in Asia [100]. Several investigations have identified that almost all ENKTL lesions are

fluorodeoxyglucose (FDG) avid [91]. In patients with ENKTL, the use of 18F-FDG positron emission tomography/computed tomography (PET/ CT) for staging is widespread [101, 102]. Nevertheless, many contradictions exist pertaining to the value of 18F-FDG PET/CT in predicting the prognosis of ENKTL [103, 104].

Some studies [105, 106] have reported that maximum standardized uptake value (SUVmax) of pretreatment 18F-FDG PET/CT is not a statistically significant predictor of overall survival and progression-free surviva (PFS). Tumor 18F-FDG uptake cannot reflect the aggressive biologic behavior of ENKTL; however, some studies have reported contradictory results [107]. These studies found that high tumor 18F-FDG uptake was closely associated with unfavorable treatment and survival outcomes. Chang et al. [108] reported that baseline wholebody total lesion glycolysis (TLG) was a good predictor of PFS and overall survival in patients with ENKTL. However, treatment plans were not uniform in these studies, potentially affecting the treatment outcome and predictive value of pretreatment 18F-FDG PET/CT. Prospective research methods have also been used to assess the prognostic value of 18F-FDG PET/CT in ENKTL [109], but considering some uncertainty in the reported results, it remains unclear. A novel solution is accordingly needed. Although deep learning has been advantageous in assisting molecular imaging to optimize therapeutic prognosis [53], it is extremely difficult to develop appropriate deep learning methods for this rare condition with only a limited number of cases.

We herein propose a weakly supervised deep learning (WSDL) method based on positive–negative unlabeled (PNU) classification [110] to maximize the utility of incomplete and missing follow-up data so as to predict the prognosis of ENKTL. We investigated the accuracy and robustness of this data enhancement strategy on a retrospective cohort to test a therapeutic regime for ENKTL. The algorithm for the WSDL method is summarized as follows:

- Train deep convolutional neural networks (DCNNs) with labeled data to obtain the baseline model.
- Use baseline DCNNs to extract features from labeled and unlabeled data.
- Build the PNU classifier to generate implicit labels for unlabeled data.
- Re-train DCNNs with labeled and unlabeled data to obtain the final prognosis.

Chapter 6: Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features

Artifacts, blur, and noise are the common distortions degrading MRI images during the acquisition process, and deep neural networks have been demonstrated to help in improving image quality. Besides, graph Neural Networks (GNNs) [111] have also shown their powerful ability to exploit structural information dealing with data of graph structure. The notation of GNNs was firstly introduced, and then further elaborated as a generalization of recursive neural networks, which is widely used to explore the structural characters in various applications, including chemistry, recommender systems, and social network study to deal with challenge tasks, e.g., finding the chemical compounds that are most similar to a query compound, tackling the graph similarity computation for query systems [112].

Nowadays, it is an interesting trend to combine GNN and CNN to develop their corresponding advantages [113]. GNNs help with reducing the data dimensionality from image features extracted by CNN to high-level and compact features in graph nodes. FCNs are limited in the receptive field. Adding GNNs could increase the receptive field of networks when dealing with large images. The combination of CCN and GNN is a convolutional graph neural network that generalizes the operation of convolution from grid data to graph data. It plays a central role in building up many complex GNN models [114]. Motivated by the idea that to learn the symmetry and self-similarity relationship of patch-based features in multi-modal brain MR images where the structure of the brain is normally symmetry, we designed a specialized Graph-based structure to merge the high-similarity information of sub-regions by updating larger weights to the more important and similar nodes or features in a graph attention fashion. Specifically, We propose a Feedback Graph Attention Convolutional Network (FB-GACN) for MR image enhancement using a self-similarity learning strategy to update the features of each node in a graph. Learning the symmetry and similarity relationship of each pair, the content with same texture (e.g., edges, corners, and lesions) gets sharper and can be used to remove some artifacts. It recovers more texture details by employing the feedback mechanism (consecutive iterations) to facilitate low-resolution images to reconstruct super-resolution images. We demonstrate the performance in two crucial tasks: i) cross-protocol super resolution of diffusion MRI and ii) MRI artifacts removal. The proposed network achieves better high-resolution criteria and superior visual quality compared to state-of-the-art methods.

1.5 Organization

This is a publication-based thesis with the following structure: Chapter 1 introduces the topics of deep learning, medical image segmentation, prognostic diagnosis, and medical image synthesis as well as their corresponding challenges, and meanwhile, we also summarize our motivations and contributions for these tasks. In Chapter 2, We give a brief and indispensable introduction on relevant terminology or knowledge of convolutional neural networks and graph neural networks.

Chapter 3 to 6 are composed of four publications [115, 116, 117, 118], which have been published as peer-reviewed journals or conference proceedings and are therefore self-contained. Each of these chapters starts with a brief synopsis introducing the main content of the corresponding publication and a statement of author's contributions.

Chapter 7 provides discussions and conclusions and suggests some interesting relevant directions as the outlook. Finally, a complete list of publications that wrote in the period of this doctoral thesis can be found in Appendix A.

2

Background

To better follow this thesis, the current chapter introduces some basic concepts and knowledge of multi-layer perceptron (MLP), convolutional neural network (CNN), and generative Adversarial Networks (GANs). For a detailed and complete overview of state-of-the-art network architectures, please refer to these works [3, 29, 119].

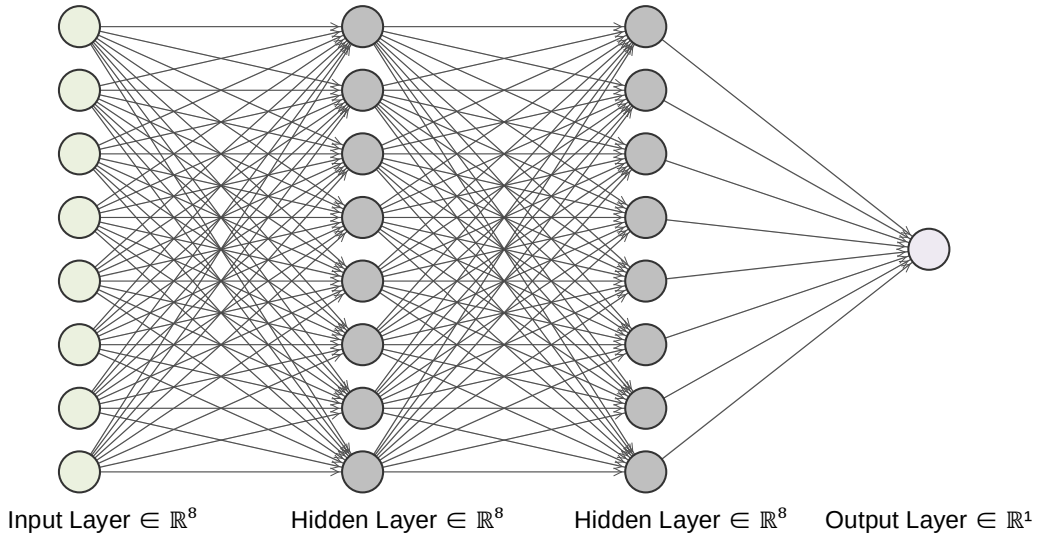


Figure 2.1: A diagram of typical multi-layer perceptron (MLP) architecture consisting of an input layer, hidden layers, and an output layer.

2.1 Feedforward Neural Network

A feedforward neural network is the simplest artificial neural network wherein connections between the nodes are feedforward without any cycles or loops. Multi-layer

perceptron is a kind of feedforward neural network. As shown in Fig. 2.1, a typical feedforward neural network or multi-layer perceptron consists of an input layer, some hidden layers, and an output layer. Each node mimics brain neurons via using a nonlinear activation function as follows:

$$\hat{f}(\mathbf{x}) = h(\mathbf{w}^{\mathbf{T}}\mathbf{x} + b) \tag{2.1}$$

where $\mathbf{x}$ means input data (e.g., images or vectors), $h(\cdot)$ is an activation function distinguishing non-linear data. $\mathbf{w} = (w_1, \cdots, w_n)$ and b are the weight and the bias vector of a hidden layer. For most tasks, the commonly existing activation functions [6] are listed as:

(1) Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

(2) Softmax

$$\sigma(x) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}} \tag{2.3}$$

(3) Rectified Linear Unit (ReLU)

$$\sigma(x) = \max(0, x) \tag{2.4}$$

(4) Leaky ReLU

$$\sigma(x) = \max(0.01x, x) \tag{2.5}$$

(5) Tanh

$$\sigma(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.6}$$

(6) Exponentional Linear Unit (ELU)

$$\sigma(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^{x} - 1), & x < 0 \end{cases} \tag{2.7}$$

To increase model nonlinear representation ability, more hidden layers can be added to the feedforward neural network. Then a feedforward neural network with a number of hidden layers can be expressed as:

$$\begin{aligned} \hat{f}(\mathbf{x}; \mathbf{\Theta}) &= (f_m \circ \cdots \circ f_1)(\mathbf{x}) \\ &= h^m \left(h^{m-1} \left(\cdots \left(h^2 \left(h^1(\mathbf{w}_1^{\mathbf{T}}\mathbf{x} + b_1\right) + b_2\right) + b_{m-1}\right) + b_m\right) \end{aligned} \tag{2.8}$$

where $\Theta = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, b_1, \cdots, b_m\}$ is the training parameter set and updated by back-propagation gradient descent strategy [120] according to the loss function representing the error between the prediction and the ground-truth. The loss function is designed depending on the practical applications, such as mean absolute error, cross-entropy loss, or dice loss.

2.2 Convolutional Neural Networks

Convolutional neural networks were designed from the inspiration of biological processes where the organization connections of the animal visual cortex were similar to the connectivity pattern between neurons. Each cortical neuron makes a response for stimulus in a restricted region of the visual field, also named the receptive field. The receptive field of each neuron is partial overlap and covers the entire visual field. Considering the vectorization of image data ruined the structural information among neighboring pixels, the convolutional neural network is proposed to tackle image structural data, which is the dominant data type in the computer vision community. There exist some representative neural architectures, ResNet [121], DenseNet [122], VGGNet [123], and Inception net [124]] for image classification or diagnosis; Fully convolutional network (FCN) [28], U-Net [33], V-Net [35] and their variants [36, 37] for image segmentation. As shown in Fig. 2.2 and Fig. 2.3, convolutional neural network commonly consists of convolutional layer, pooling layer, batch normalization, and skip connection.

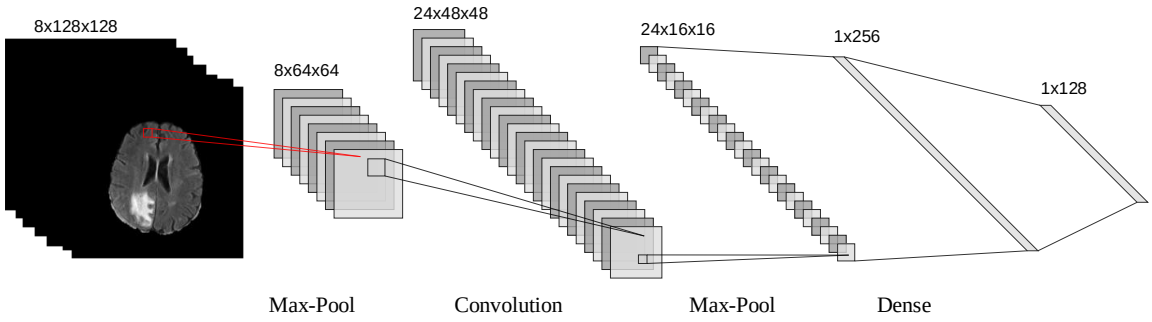


Figure 2.2: A sample convolutional network architecture mainly consisting of convolutional layers, pooling layers, and dense layers.

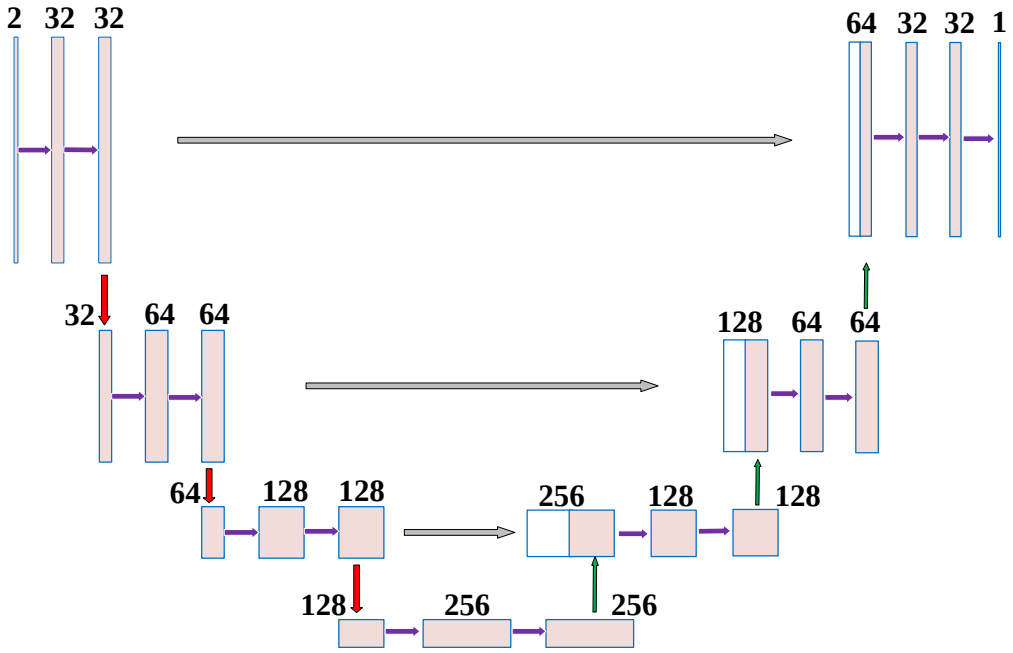


Figure 2.3: A typical convolutional network architecture for segmentation task: U-Net. The number above the feature maps means the number of channels.

2.2.1 Convolutional layer

As the core building block of a convolutional neural network, the role of the convolutional layer aims to detect local features at different positions of feature inputs from the previous layer with learnable kernels $K^l_{i,j}$. Essentially, the kernels are the connection weights between feature maps of the current layer and that of the previous layer. The convolutional process can be expressed as:

$$Y_i^{(l)} = h\left(\sum_{j=1}^{N^{(l-1)}} K_{i,j}^{(l)} * Y_j^{(l-1)} + B_i^{(l)}\right) \tag{2.9}$$

where $Y_i^{(l)}$ is the i^{th} feature map at l^{th} layer, B_i^l is a bias matrix, h is a activation function. To efficiently extract features and decrease the trainable parameters size, deformable convolution [125], depth-wise separable convolution [126], and dilated convolution [127] are proposed to improve representation ability.

2.2.2 Pooling layer

The pooling layer aims to downsample the feature maps and reduce the training parameters in the network, but meanwhile increases the receptive field of networks. Specifically, each small region is pooled by pooling operations to generate a single number that represents the information of the current small region. The pooling operations are usually chosen from the max function or the average function. Similarly, the convolutions with increased stride lengths can also get the same effect of pooling operation [4].

2.2.3 Batch normalization

Batch normalization is usually located after the activation function to normalize the feature map by subtracting the mean and dividing by standard deviation. This process is acted as the regularize of a network to speed up the training stage and make it less sensitive for parameter initialization [128].

2.2.4 Skip connection

Skip connection is a kind of residual learning to allow gradients to flow through a network without passing through intermediate operations. The gradient of a network vanishes when the depth of model increases. The skip connection can alleviate the gradient vanishment by propagating the gradient through network [121, 122]. Besides, the skip connection keeps the residual information with higher spatial priors to further complementing deep latency features.

2.2.5 Loss function

The loss function acts as an objective function role for optimization problems to supervise the training stage of networks. Thus, the loss functions are designed based on specific tasks (e.g., classification or segmentation). In the image classification task, the cross-entropy loss [129] is used to categorize input images into different classes as follows:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\delta(y_i = c)\log(P(y_i = c)) \tag{2.10}$$

where N means the N^{th} data, C is the C^{th} class, $\delta(y_i = c)$ is the indicator function and $P(y_i = c)$ is the probability belong to current C^{th} class.

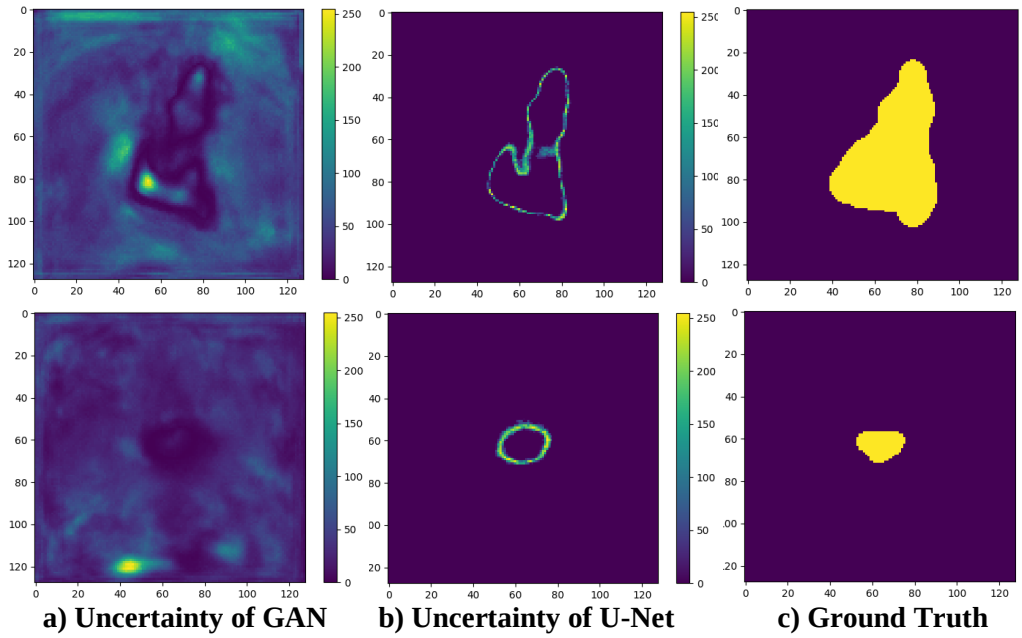


Figure 2.4: Qualitative segmentation uncertainty analysis for (a) generative adversarial networks method and (b) the U-Net, while (c) shows the corresponding ground truth. The pixel-wise uncertainty is normalized to the interval [0, 255]. Brighter zones indicate a higher uncertainty character.

For medical image segmentation task, the loss function is based on the dice score and is defined as [35]:

$$minL_{Dice} = \frac{2\sum_{n=1}^{N} p_i g_i + s}{\sum_{n=1}^{N} p_i^2 + \sum_{n=1}^{N} g_i^2 + s} \tag{2.11}$$

where g_i and p_i represent the ground-truth and predicted probabilistic pixel, respectively. The rest term s ensures stability by avoiding the division by 0. We set s to 1 in our experiments, where the entries of g and p are all zeros.

2.3 Uncertainty Analysis

Dropout is a way of Bayesian approximation. During training, the input channels x and the corresponding ground truth lesion labels Y are used to learn the weights θ of the network. To capture the uncertainty character in the model, a prior distribution is

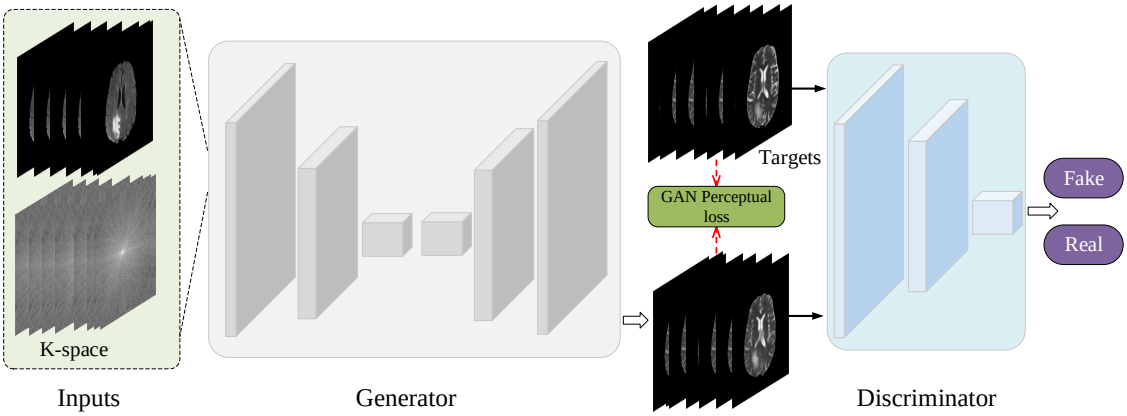


Figure 2.5: A sample of generative adversarial networks consisting of a generator and a discriminator for brain MRI synthesis.

placed over θ and an estimate of the posterior $p(\theta|X,Y)$ is calculated. An analytical computation of this prior is intractable, but variational methods can approximate it with a parameterized distribution $q(\theta)$ by minimizing the Kullback-Leibler (KL) divergence [130]:

$$q^*(\theta) = argminKL(q(\theta)||p((\theta|X,Y))_{q(\theta)} \tag{2.12}$$

According to [96], Yarin et al. declare that minimizing the cross-entropy loss of a network with dropout applied after each layer of weights is equivalent to the minimization of the KL-divergence. In order to analyze the reliable capability of the network, the Monte Carlo dropout method is introduced here. Additionally, a novel corresponding uncertainty evaluation criterion is proposed to measure the network's resistance to the epistemic uncertainty according to the variance map, which is directly obtained from the probability map. The variance map intuitively reflects the prediction fluctuation of the network architecture. The uncertainty sources are mainly from two aspects, background and lesion. As shown in Fig. 2.4, the uncertainty from the lesion-zone causes the lesion prediction error and severely affects many important criteria such as the Dice Score and Sensitivity whose evaluation policy relies on the lesion region and its boundary pixels. In sharp contrast to the uncertainty from the lesions, the uncertainty from the background mostly affects the criteria, which rely on pixels predicted as non-lesion pixels, such as specificity (true negative rate), which measures the proportion of actual negatives that are correctly identified.

2.4 Generative Adversarial Networks

Adversarial learning has gained plenty of attention from industry and academic community due to its effectiveness in tackling domain shift and generating image samples [131]. Basically, GAN [63] is a kind of neural network architecture where two networks are trained simultaneously, a generator aims to generate images close to ground truth, and a discriminator to distinguish between fake or real images.

For the generator G, the input is random noise sampled from a distribution usually chosen from a uniform or Gaussian. The output of the generator is encouraged to be closer to real samples from real data distribution. Fig. 2.5 shows an illustration of GAN architecture and generator of this example is to synthesize brain MRI images. For the discriminator D, the input is either a real sample or a fake sample and the corresponding output is an indicator showing the probability of input being a real or fake sample. The objective of generator G and discriminator D can be formulated as [131]:

$$\mathcal{L}_D^{GAN} = \max_{\theta_D} \mathbf{E}_{\mathbf{x}\sim\mathbf{P_{data}}}[\log \mathbf{D}(\mathbf{x})] + \mathbf{E}_{\mathbf{z}\sim\mathbf{P_z}}\log(\mathbf{1} - \mathbf{D}(\mathbf{G}(\mathbf{z}))) \tag{2.13}$$

$$\mathcal{L}_G^{GAN} = \min_{\theta_G} \mathbf{E}_{\mathbf{z}\sim\mathbf{P_z}}\log(\mathbf{1} - \mathbf{D}(\mathbf{G}(\mathbf{z}))) \tag{2.14}$$

where x is a real image from the unknown data distribution P_{data}, and z is a random input for the generator, following a probability distribution. θ_G and θ_D represent the parameters for the generator and the discriminator in a GAN. The ideal outcome after training is that the samples synthesized by a generator approximate to the real data distribution.

3

Hierarchical Multi-class Segmentation of Glioma Images Using Networks with Multi-level Activation Function

This chapter has been published as **peer-reviewed conference paper:**.

X. Hu, H. Li, Y. Zhao, C. Dong, B. H. Menze, and M. Piraud. "Hierarchical multi-class segmentation of glioma images using networks with multi-level activation function." In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 116–127

Synopsis: This work deals with the problem of multi-class segmentation of Glioma images. We present a multi-level activation function incorporating nesting topological priors for MICCAI Brain tumor segmentation challenge with its three hierarchically nested classes 'whole tumor', 'tumor core', 'active tumor'. Based on the same start-of-the-art network architecture, the accuracy of nested-class (enhancing tumor) is reasonably improved from 69% to 72% compared with the traditional Softmax-based method which blind to topological prior.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

Accepted manuscript for the International MICCAI Brainlesion Workshop, Springer
Published version: https://doi.org/10.1007/978-3-030-11726-9_11

Hierarchical Multi-class Segmentation of Glioma Images Using Networks with Multi-level Activation Function

Xiaobin Hu[1,2], Hongwei Li[1], Yu Zhao[1], Chao Dong[1], Bjoern H. Menze[1], and Marie Piraud[1]

[1] Department of computer science, Technische Universität München, Munich, Germany
[2] xiaobin.hu@tum.de

Abstract. For many segmentation tasks, especially for the biomedical image, the topological prior is vital information which is useful to exploit. The containment/nesting is a typical inter-class geometric relationship. In the MICCAI Brain tumor segmentation challenge, with its three hierarchically nested classes 'whole tumor', 'tumor core', 'active tumor', the nested classes relationship is introduced into the 3D-residual-Unet architecture. The network comprises a context aggregation pathway and a localization pathway, which encodes increasingly abstract representation of the input as going deeper into the network, and then recombines these representations with shallower features to precisely localize the interest domain via a localization path. The nested-class-prior is combined by proposing the multi-class activation function and its corresponding loss function. The model is trained on the training dataset of Brats2018, and 20% of the dataset is regarded as the validation dataset to determine parameters. When the parameters are fixed, we retrain the model on the whole training dataset. The performance achieved on the validation leaderboard is 86%, 77% and 72% Dice scores for the whole tumor, enhancing tumor and tumor core classes without relying on ensembles or complicated post-processing steps. Based on the same start-of-the-art network architecture, the accuracy of nested-class (enhancing tumor) is reasonably improved from 69% to 72% compared with the traditional Softmax-based method which blind to topological prior.

Keywords: Topological prior · nested classes · 3D-residual-Unet · multi-class activation function

1 Introduction

Glioma are the most common family of brain tumors, and forms some of highest-mortality and economically costly diseases of brain cancer [1–3]. The diagnosed method is highly relayed on manual segmentation and analysis of multi-modal MRI scans by bio-medical experts. Nevertheless, this diagnosed way is severely limited by the labor-intensive character of the manual segmentation

process and disagreement or mistakes between manual segmentation. Consequently, there exists a great need for a fast and robust automated segmentation algorithm. Convolutional neural networks (CNNs) have been verified to be extremely effective for a variety of semantic segmentation tasks [4].

While CNN segmentation algorithms are abundant in biomedical imaging, only very few make use of nested-topological prior information. Among the few that do [5–11], we find three different approaches. First, the use of cascaded algorithms where the network consists of successive segmentation networks. Second, the information on the nested-classes is incorporated into the loss function, imposing penalties on solutions that do not respect the nested geometry relations. Third, Markov random fields are used to formalizing class relationship in the post-processing of the network output. Here, we make use of a new activation function [12] that is directly implementing class hierarchy in the network training and generalize it to 3 nested classes. For the glioma labels we assume that active tumor regions are always contained in the tumor core which is surrounded by the tumor edema, resulting in a hierarchical three-class model. In sharp contrast with nested-class method, the softmax-based method of multi-class ignores the geometric prior between different classes, and assumes the classes are mutually-exclusive, meaning one pixel cannot belong to different classes at the same time, which absolutely discards the topological information and sometimes leads the unreasonable segmentation results. The comparison of Dice score criteria between two different methods is implemented and it obviously indicates the nested-class method achieves higher accuracy than the softmax-based method, especially for the internal-classes.

In the following, we introduce a brief overview of start-of-the-art 3D-residual U-net architecture and multi-class-nested activation and loss function. We then propose and evaluate our model architectures for Brats tumor segmentation. Finally, we implement the comparison between two main avenues and illustrate the multi-level activation performs better especially in the inter-class.

2 Methodology

2.1 Network Architecture

The nested-classes relationship between different labels are shown in Fig.2. The general network structure shown in Fig.1 is stemming from the previously used glioma segmentation network by Isensee [13] to process large 3D input blocks of 144x144x144 voxels. The original network is inspired by the U-net [14] which allows the network to intrinsically recombine different scales throughout the entire network. This vertical depth is set as 5, which balances between the spatial resolution and feature representations. The context module is a preactivation residual block, and is connected by 3x3x3 convolutions with input stride 2. The purpose of the localization pathway is to extract features from the lower levels of the network and transform them to a high spatial resolution by means of a simple upscale technology. The upsampled features and its corresponding level of the context aggregation feature are recombined via concate-

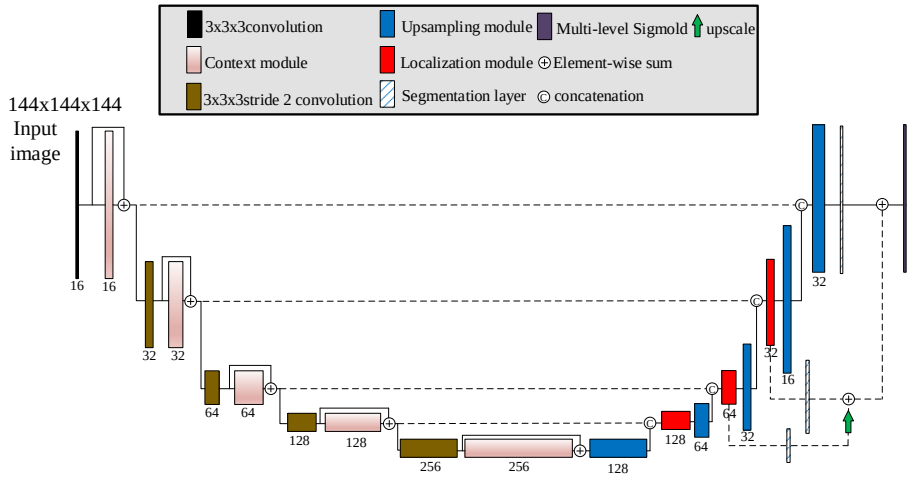


Fig. 1. Network architecture from [13]: Context pathway (left) aggregates high level information;Localization pathway (right) localizes precisely

nation. Furthermore, the localization module, consisting of a 3x3x3 convolution followed by a 1x1x1 convolution, is designed to gather these features.

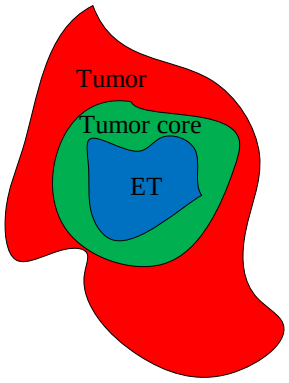


Fig. 2. Schematic description of the nesting of classes in the BRATS challenge, which respects the following hierarchy: Enhancing Tumor (ET) ∈Tumor core ∈ Tumor

The deep supervision is introduced in the localization pathway by integrating segmentation layers at different levels of the network and combining them via

elementwise summation to form the final network output. The output activation layer is multi-level Sigmoid layer instead of softmax layer in the Isensee's network which converting the multi-class problem to binary ones. Intrinsically, the multi-level activation is the assemble of multi-sigmoid function and then straightforwardly maps to multi-class segmentation incorporating the topological prior. Consequently, it overcomes the softmax-based method's shortcoming which is blind to the geometric prior.

2.2 Crop preprocessing

For 3D network architecture, the larger patch size of training dataset contains more continuous context knowledge and localization information which are beneficial to improve the segmentation accuracy. In order to acquire to the larger cube size patch of 3D image, the valuable knowledge in the MRI is extracted as much as possible while the meaningless information is cropped. Then the crop processing is implemented, and the maximum size of cube patch is selected as [144,144,144].

The crop preprocessing equation is defined as:

$$\begin{aligned} array &= [a_{min} - (b_{size} - a)/2 : a_{min} + (b_{size} + a)/2] \\ a &= a_{max} - a_{min} \end{aligned} \tag{1}$$

where a_{min} and a_{max} are the min and max non-zero information index of MRI image, and a represents the length of non-zero information.b_{size} is the cube patch size and selected as 144.

The index is recorded and used in the image post-processing stage to recovery back to the original shape [155,240,240]. However, a little of meaningful information which exceeds the cube patch size 144 is unavoidably ignored and have little effect on the segmentation result. In order to equally compare the softmax-based with the multi-level method, no data augmentation operation is used in the stage of image pre-procssing.

2.3 Multi-level method

Here, we use one output channel and a multi-class-nested activation function, as first proposed in [12].The multi-level method is inspired by continuous regression, and thereby generalizing logistic regression to hierarchically-nested classes. It is shown in Fig.3 and defined as

$$a(x) = \sum_{n=1}^{m} \sigma(k[x + h(n - \frac{m+1}{2})]) \tag{2}$$

Where σ is the sigmoid function, k is the steepness and h is the spacing between consecutive Sigmoids. For Brain tumor segmentation challenge 4-classes nested label case, we have m+1=4, and we take h=0.5 and steepness=10. The

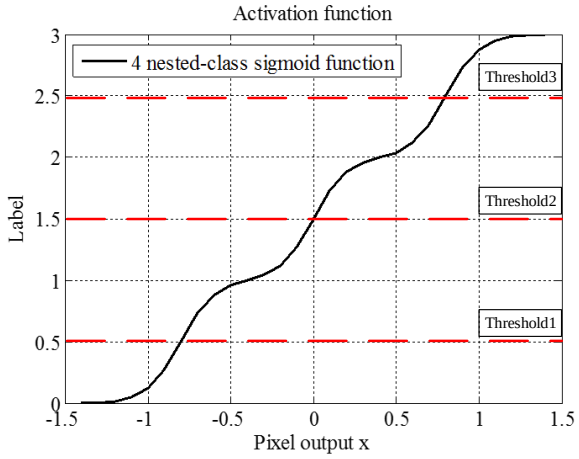


Fig. 3. Multi-class activation function, Eq.(1) with m+1=4, h=0.8 and k=10

corresponding loss function, called Modified Cross-Entropy (MCE) in [12], is defined as

$$L_{MCE} = -\frac{1}{N_{tot}} \sum_{pixel\ i} \sum_{classes\ c} y_i^c w^c log(P^c[a(x_i)]) \tag{3}$$

where w^c is the weight of corresponding label,which we take as $w^{c\alpha}(w^{c\alpha} = (\frac{N_{tot}}{N_c})^\alpha)$, where $N_{tot}$is the sum number of pixels,N_c the number of pixels in each class, and where $y^c = 1$ for the ground-truth label c of pixel i and $y^c = 0$ otherwise. Furthermore, the mapping function P^c is defined as

$$\begin{aligned} P^{c=0}(a) &= 1 - a/3 \\ P^{c=1}(a) &= a\Theta(1-a) + (3-a)/2\Theta(a-1) \\ P^{c=2}(a) &= a/2\Theta(2-a) + (3-a)\Theta(a-2) \\ P^{c=3}(a) &= a/3. \end{aligned} \tag{4}$$

Where $\Theta(x)$ is the Heaviside function. The other one loss function, called Normalized Cross-Entropy (NCE) in [12], is defined as

$$L_{NCE} = -\frac{1}{N_{tot}} \sum_{pixel\ i} \sum_{classes} y_i^c w^c log(\Theta^c[a(x_i)]) \tag{5}$$

Furthermore, the mapping function Q^c is defined as

$$\begin{aligned} Q^{c=0}(a) &= s(1-a) \\ Q^{c=1}(a) &= a\Theta(1-a) + s(2-a)\Theta(a-2) \\ P^{c=2}(a) &= s(a-1)\Theta(2-a) + (3-a)\Theta(a-2) \\ P^{c=3}(a) &= s(a-2). \end{aligned} \tag{6}$$

where s is the softplus function,, and $\Theta(x)$ is the Heaviside function.

Weighted modified and Normalized cross-entropy losses are naturally combined with standard cross-entropy loss and mitigate the class unbalance problem. They also have the ability to encode of any hierarchical and mutually-exclusive topological relationship of classes in a network architecture.

2.4 Evaluation metrics

In the task for BRATS, the number of positives and negatives are highly unbalanced. Consequently, four typical different metrics are used by the organizers to evaluate the performance of the algorithm and then rank the different teams.

Give a ground-truth segmentation map G and a segmentation map corresponding one class generated by the algorithm. The four evaluation criteria are defined as following.

Dice similarity coefficient(DSG):

$$DSC = \frac{2(G \cap P)}{|G| + |P|} \tag{7}$$

The Dice similarity coefficient measures the overlap in percentage between G and P.

Hausdorff distance (95th percentile) is defined as :

$$H(G,P) = max(supinf_{x \in G, y \in P} d(x,y), supinf_{y \in P, x \in G} d(x,y)) \tag{8}$$

where $d(x,y)$ denotes the distance of x and y, sup denotes the supremum and inf for the infimum. This measures how far two subsets of a metric space are from each other. As used in this challenge, it is modified to obtain a robustified version by using the 95th percentile instead of the maximum(100 percentile) distance.

Sensitivity (also called the true positive rate) measures the proportion of actual positives that are correctly identified. Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified. Assume P is the number of real positive prediction pixel of lesion and N is the number of real negative prediction pixel of lesion. Condition positive P consists with true positive TP and false negative FN. Besides, the condition negative N is also divided into TN true negative and FP false positive.

Then, the metrics of Sensitivity and Specificity are illustrated as:

$$Sensitivity = \frac{TR}{P} = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{10}$$

Then the values of those four metrics were computed by the organizers independently and made available in the validation leaderboard.

3 Experiment results

In BRATS 2018 dataset [15–19], there are four types, Necrotic core, Edema, Non-enhancing core and Enhancing core that form the three tumor classes in Fig.2. The dataset contains 4 different modalities for MRI, native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 Fluid Attenuated Inversion Recovery (FLAIR) which are all used as different input channels. We train the networks using ADAM optimizer with an initial learning rate of 0.0005, and to regularize the network, we use early stopping when the precision on the 20% of the training dataset reserved for validation is no longer improved, and dropout (with rate 0.3) in all residual block before the multi-class sigmoid function. Some slices of segmentation results containing the tumor, tumor core and enhancing core are shown in Fig.4. We observe that the topology geometry between different labels is constrained to the nested-classes relationship, consequently avoiding errors stemming from the lack of topological prior.

	Dice score			
	Enhancing core	whole tumor	tumor Core	Weight scheme
Multi-level(MCE)	**0.719**	0.857	**0.769**	0.4
Multi-level(NCE)	0.676	0.857	0.755	0.4
Multi-level(NCE)	0.633	0.837	0.736	0.5
Multi-level(NCE)	0.655	0.856	0.758	0.3
Softmax-based method	0.691	**0.861**	0.763	-

Table 1. Validation results presented on the leaderboard

The segmentation result is severely affected by highly unbalanced problems existing in the Brats dataset. As class imbalance in a data set increases, the performance of a neural net trained on that data has been shown to decrease dramatically [20]. In order to mitigate this issue, many methods [21–23] were proposed to modify the loss function to alleviate this problems. Here,the weighted cross entropy incorporating the nested-class information is proposed and investigated. We experimented with different weighting schemes (α=1,0.5,0.4,0.3) and with the different losses MCE and NCE proposed in [12]. The best performing combination turned out to be α=0.4 and MCE loss function. The segmentation thresholds to determine the boundaries between classes, were set to [0.95,1.65,2.2] on

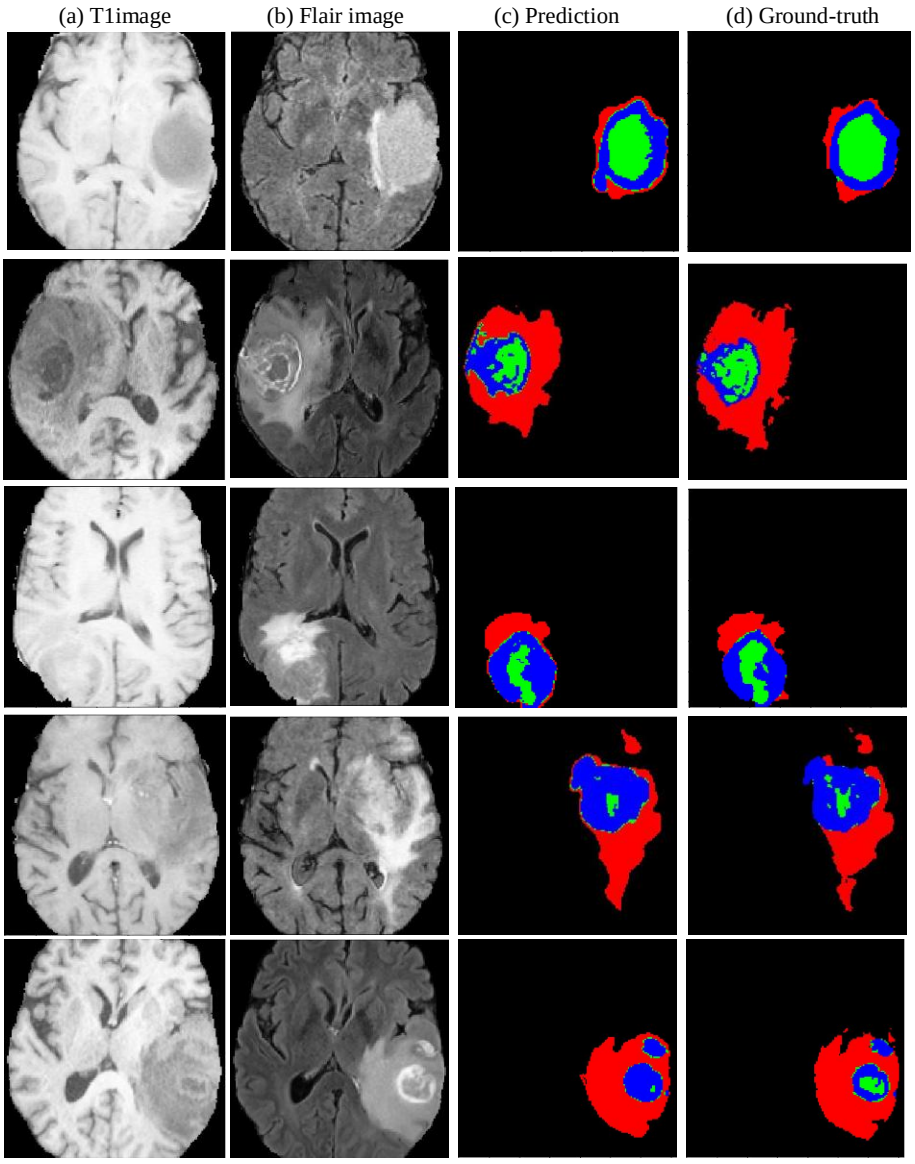


Fig. 4. Segmentation results, for five different validation cases. The tumor class is depicted in red, tumor core in green and enhancing tumor in blue.

the validation process. For this final configuration, we reached Dice scores of 86% for the complete tumor, 77% for the tumor core and 72% for the enhancing core as presented in Table 1. The weighted-modified-cross-entropy performs much better than the result achieved by normalized cross-entropy, and weight scheme affects the segmentation result severely since the extraordinary unbalance prob-

Dice score	Enhancing core	whole tumor	tumor Core
Mean	0.71965	0.85685	0.76906
StdDev	0.28526	0.09802	0.21962
Median	0.84268	0.87823	0.84325
25quantile	0.6889	0.83379	0.70743
75quantile	0.8876	0.90895	0.91292

Table 2. Quantitative evaluation of Dice score

lem. The different weight schemes $[0.5, 0.4, 0.3]$ are compared and the optimal weight scheme is taken as 0.4. In comparison with the softmax-based method based on the same network architecture proposed by Isensee without ensembles operation, any complicated image pre-processing and post-processing steps and extra training dataset, it indicates that the Dice score of nested-class (enhancing core) drastically improved from 0.691 to 0.719 while the Dice core of whole tumor and tumor core almost remains at same extent. The quantitative evalu-

Mean	Enhancing core	whole tumor	tumor Core
Sensitivity	0.74119	0.93916	0.78743
Specificity	0.9974	0.98715	0.99591
Hausdorff95	5.50007	10.84397	9.98557

Table 3. Sensitivity, Specificity and Hausdorff95 results presented on the leaderboard

ation (Mean, std, Median, 25%, 75% quantile) of Dice score of enhancing core and whole tumor and tumor core are showed in Table 2. And other evaluation metrics (the proportion of actual positives correctly identified—Sensitivity, the proportion of actual negatives correctly identified—Specificity and Hausdorff95) are listed in Table 3.

3.1 Threshold scheme definition and analysis

Setting the optimal threshold is an important component of the multi-class segmentation task, and it is straightforwardly linked to segmentation boundary. From the activation function (4 nested-class sigmoid function) Fig.3, the 4 classes segmentation problem is corresponding with the threshold scheme with 3 parameters [Threshold-1, Threshold-2, Threshold-3]. The threshold scheme is optimally chosen during the validation procedure, and then fixed and applied into test dataset.

In order to analyze how the threshold affects the segmentation accuracy, the relationship between boundary threshold and Dice score is illustrated in Fig.5. The target threshold is changed to the value taken from a specific interval which is considered to be possible to achieve optimal segmentation result when other

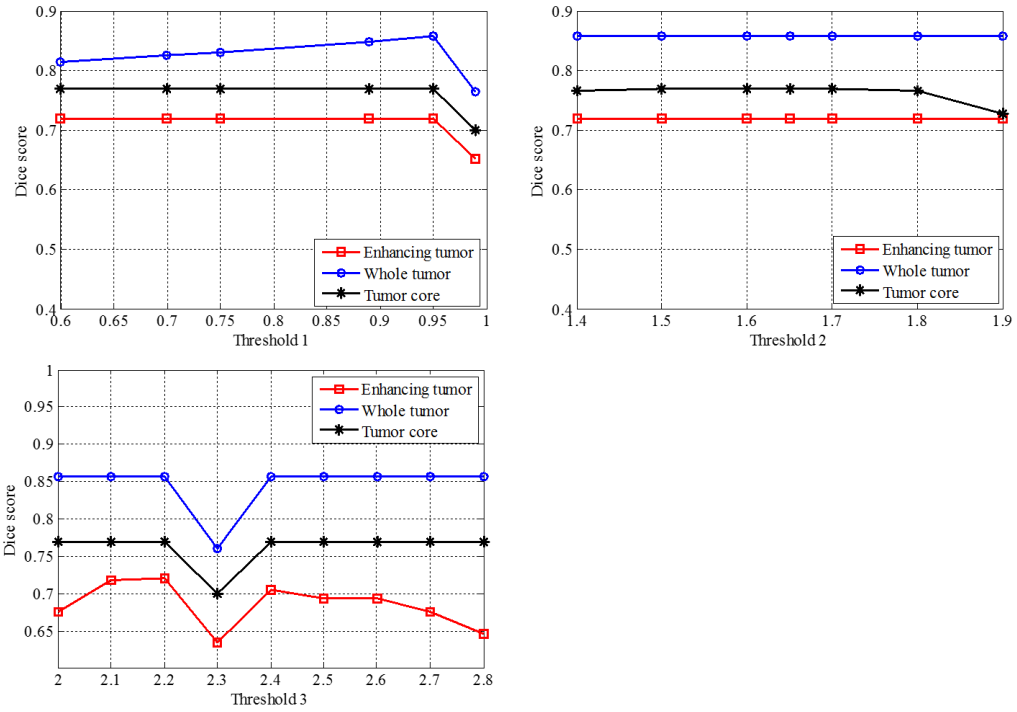


Fig. 5. Boundary division of Threshold scheme

thresholds are fixed at the optimal value. The criteria Dice score of three classes is very sensitive to the threshold-3 value compared with other two threshold indexes, that it may drop into Dice score valley within interval [2.2,2.4]. The threshold-2 index has little impact on the Dice score of whole classes except for threshold greater than 1.8. Consequently, it is easier to make an optimal threshold scheme after determining indexes of threshold-3 and threshold-2. After experiment and optimization, the suitable threshold scheme in the Brats challenge is selected as [0.95,1.65,2.2].

4 Conclusions

In this paper we applied the technique of multi-level activation to the nested classes segmentation of glioma. The results of our experiments indicate that the multi-level activation function and its corresponding loss function are efficient compared to Softmax output layer based on the same network framework. Using the MCE loss function and a reweighting scheme with power-law =0.4, we obtain Dice scores 86% for complete tumor, 77% for tumor core and 72% for enhancing core on the validation leaderboard of the 2018 BRATS challenge, proving the

applicability of the multi-level activation scheme. Finally, this activation could be combined with other network architectures. Using it with the best performing architecture of the BRATS challenge could even lead to further improved results.

References

1. Davis M.E.: Glioblastoma: Overview of Disease and Treatment. Clinical journal of oncology nursing. **20**(5),S2-S8(2016) https://doi.org/10.1188/16.CJON.S1.2-8
2. Hanif, F., Muzaffar, K., Perveen,K., Malhi, S.M., Simjee,S.U.: Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pacific Journal of Cancer Prevention. **18**,3–9 (2017)
3. Birbrair, A., Sattiraju, A., Zhu, D., Zulato, G., Batista, I., Nguyen, V.T., Messi, M.L., Solingapuram, Sai.K.K., Marini, F.C., Delbono, O., Mintz, A.: Novel Peripherally Derived Neural-Like Stem Cells as Therapeutic Carriers for Treating Glioblastomas. STEM CELLS Translational Medicine. **6**,471–481 (2017)
4. Gu,J.X., Wang,Z.H., Kuen,J., Ma, L.Y., Shahroudy, A., Shuai,B., Liu, T., Wang, X.X., Wang.L., Wang,G., Cai, J.F., Chen, T.: Recent Advances in Convolutional Neural Networks. Pattern Recognition. **77**,354–377 (2018)
5. Nosrati, M.S., Hamarneh, G.: Local optimization based segmentation of spatially-recurring, multi-region objects with part configuration constraints. IEEE Transactions on Medical Imaging **33**, 1845–1859 (2014)
6. BenTaieb, A., Hamarneh, G.: Topology Aware Fully Convolutional Networks for Histology Gland Segmentation. In Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W., eds.: MICCAI 2016. Lecture Notes in Computer Science, Cham, Springer International Publishing (2016)
7. Christ, P.F., Elshaer,M.E.A., Ettlinger,F., Tatavarty,S., and Bickel, M., Bilic,P., Rempfler, M., Armbruster, M., Hofmann, F., Anastasi, M.D.,Sommer,W.H.,Ahmadi,S.A.,Menze,B.H.: Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields. In Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W., eds.: MICCAI.Lecture Notes in Computer Science **9900**, (2016)
8. Fidon, L., Li,W.Q., Garcia-Peraza-Herrera,L.C., Ekanayake,J., Kitchen,N., Ourselin,S., Vercauteren,T.: Generalised Wasserstein Dice Score for Imbalanced Multi-class Segmentation using Holistic Convolutional Networks. Oral presentation at the MICCAI 2017 Brain Lesion (BrainLes) Workshop 1-12, (2014)
9. Bauer, S., Tessier,J., Krieter, O., Nolte, L.P., Reyes,M.: Integrated spatio-temporal segmentation of longitudinal brain tumor imaging studies. In Menze, B., Langs, G., Montillo,A., Kelm, M., Müller, H., Tu, Z., eds.: Medical Computer Vision. Large Data in Medical Imaging, Cham, Springer International Publishing 71-83, (2014)
10. Alberts, E., Charpiat,G., Tarabalka,Y., Huber,T., Weber, M.A., Bauer,J., Zimmer,C., Menze,B.H.: A Nonparametric Growth Model for Brain Tumor Segmentation in Longitudinal MR Sequences. MICCAI Brain Lesion Workshop 69-79, (2015)
11. Liu, Z.W., Li,X.X., Luo,P., Loy, C.C., Tang, X.O.: Deep Learning Markov Random Field for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 1-1,8828 (2017)
12. Piraud, M., Sekuboyina,A., Menze,B.H.: Multi-level Activation for Segmentation of Hierarchically-nested Classes. Computer Vision and Pattern Recognition workshop (2018)

13. Isensee, F., Kickingereder.P., Wick,W., Bendszus,M., Maier-Hein,K.H.: Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. MICCAI BraTs Challenge, (2017)
14. Ronneberger,O., Fischer,P., Brox,T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS 234-241, (2015)
15. Menze B.H., Jakab A., Bauer S., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34(10),1993-2024(2015)
16. Bakas S., Akbari H., Sotiras A., et al..: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Nature Scientific Data, (2017)
17. Bakas S., Akbari H., Sotiras A., Bilello M., Rozycki M., Kirby J., Freymann J., Farahani K., Davatzikos C.: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
18. Bakas S., Akbari H., Sotiras A., Bilello M., Rozycki M., Kirby J., Freymann J., Farahani K., Davatzikos C.: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. The Cancer Imaging Archive (2014). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
19. Spyridon Bakas, Mauricio Reyes, et Int, and Bjoern Menze, Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv preprint arXiv:1811.02629, 2018
20. Mazurowski M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks **21**(2),427–436 (2017)
21. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Fourth International Conference on 3D Vision **16**,565–571 (2016)
22. Sudre, C.H., Li, W.Q., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso M. et al. (eds) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2017, ML-CDS 2017. Lecture Notes in Computer Science **10553**, (2017)
23. Crum, W.R., Camara, O., Hill, D.L.G.: Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. IEEE Transactions on Medical Imaging **25**(11),1451–1461 (2006)

4

Coarse-to-Fine Adversarial Networks and Zone-Based Uncertainty Analysis for NK/T-Cell Lymphoma Segmentation in CT/PET Images

This chapter has been published as **peer-reviewed journal paper:**.

X. Hu, R. Guo, J. Chen, H. Li, D. Waldmannstetter, Y. Zhao, B. Li, K. Shi, and B. Menze. "Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images." In: *IEEE journal of biomedical and health informatics* 24.9 (2020), pp. 2599–2608

Synopsis: This work proposed a coarse-to-fine adversarial network for ENKTL semantic segmentation. The coarse stage acts as a dimensionality reduction to roughly locate the lesion bounding box and crop redundant information to facilitate the fine segmentation. The fine segmentation is an end-to-end adversarial network with a generator and a discriminator part. An exploration of zone-based uncertainty estimates based on Monte Carlo (MC) dropout was presented for segmentation networks. Uncertainty analysis makes it possible to permit subsequent optimization by engineers and revision by clinicians.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

Coarse-to-Fine Adversarial Networks and Zone-based Uncertainty Analysis for NK/T-cell Lymphoma Segmentation in CT/PET images

Xiaobin Hu*, Rui Guo*, Jieneng Chen, Hongwei Li, Diana Waldmannstetter, Yu Zhao†, Biao Li†, Kuangyu Shi and Bjoern Menze

***Abstract*—Extranodal natural killer/T cell lymphoma (ENKL), nasal type is a kind of rare disease with a low survival rate that primarily affects Asian and South American populations. Segmentation of ENKL lesions is crucial for clinical decision support and treatment planning. This paper is the first study on computer-aided diagnosis systems for the ENKL segmentation problem. We propose an automatic, coarse-to-fine approach for ENKL segmentation using adversarial networks. In the coarse stage, we extract the region of interest bounding the lesions utilizing a segmentation neural network. In the fine stage, we use an adversarial segmentation network and further introduce a multi-scale L_1 loss function to drive the network to learn both global and local features. The generator and discriminator are alternately trained by backpropagation in an adversarial fashion in a min-max game. Furthermore, we present the first exploration of zone-based uncertainty estimates based on Monte Carlo dropout technique in the context of deep networks for medical image segmentation. Specifically, we propose the uncertainty criteria based on the lesion and the background, and then linearly normalize them to a specific interval. This is not only the crucial criterion for evaluating the superiority of the algorithm, but also permits subsequent optimization by engineers and revision by clinicians after quantitatively understanding the main source of uncertainty from the background or the lesion zone. Experimental results demonstrate that the proposed method is more effective and lesion-zone stable than state-of-the-art deep-learning based segmentation model.**

***Index Terms*—coarse-to-fine adversarial network, multi-zone uncertainty estimate, medical image segmentation, Monte Carlo dropout**

I. Introduction

Extranodal natural killer/T cell lymphoma, nasal type (ENKL) is a rare disease, which is much more prevalent in Asia and South America. ENKL occurs predominantly in the nasal, paranasal and oropharyngeal sites. 18F-FDG PET/CT scanning is currently the most effective imaging modality for staging, monitoring response, and predicting prognosis for many kinds of lymphomas [1]. Several investigations identified that almost all ENKL are FDG avid [2], [3]. ENKL segmentation results provide the vital priors of the disease, such as lesion location and the lesion volume size. According to the lesions priors and other diagnose information, the diseases are evaluated and cataloged into different stages. For each stage, there exists a corresponding suitable treatment planning. Furthermore, the data fusion based on the lesion priors from segmentation and others (e.g. patients age, surgery performance, recovery after surgery) is used to predicate the survival time and the risk of the disease recurrence. Consequently, judging the scope of ENKL violation is very important for the patients' staging and prognosis. But the current method is highly dependent on the manual segmentation and the analysis of multi-modal scans by bio-medical experts. Moreover, diagnosis like this is severely limited by the labor-intensive character of the manual segmentation process and mistakes in manual segmentations. Consequently, there exists a great need for a fast and robust automated segmentation algorithm [4]–[8]. In this paper, we focus on lymphoma segmentation from CT and PET scans.

Convolutional neural networks (CNNs) have been verified to be extremely effective for a variety of semantic segmentation tasks [9]–[12]. For pixel-wise semantic segmentation, CNNs have also achieved remarkable successes. The first fully convolutional network (FCN) was proposed by Long et al. for semantic segmentation [13]. Here, the fully connected layers in CNNs are replaced by convolutional layers and a skip architecture is defined to combine semantic information from a shallow, fine layer to produce accurate and detailed segmentations. Noh et al. [14] proposed an encoder-decoder structure for a semantic segmentation algorithm using a deep deconvolution network. Conditional Random Fields (CRFs) and CNNs are combined for a better exploration of spatial correlations between pixels by Lin et al. [15]. Ronneberger et al. [16] presented a more elegant architecture, the U-Net. This network consists of a contracting path to capture context and a symmetric expanding path that enables precise localization, for segmentation of neuronal structures in electron microscopic stacks. Inspired by the idea of skip-connection [17], the U-Net architecture improves its performance using a large margin, and has been successfully applied and modified into different tasks [17]. Havarei [18] explored a cascade architecture where two CNNs are concatenated to gain additional information, achieving good performance for brain tumor segmentation.

Xiaobin Hu is with the Department of Computer Science, Technische Universität München, Munich, Germany e-mail:(xiaobin.hu@tum.de).
Hongwei Li, Yu Zhao, Diana Waldmannstetter and Bjoern Menze are with the Department of Computer Science, Technische Universität München, Munich, Germany e-mail: (bjoern.menze@tum.de).
Jieneng Chen is with the College of Electronics and Information Engineering, Tongji University, Shanghai, China.
Rui Guo and Biao Li is with the Department of Nuclear Medicine, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine.
Kuangyu Shi is with the Lab for Artificial Intelligence Translational Theranostics, Dept. Nuclear Medicine, University of Bern.
The authors marked with * make the equal contribution to the paper.
The corresponding author is marked with †.

Kamnitsas et al. [19] proposed a 3D CNN using two pathways with inputs of different resolutions. Additionally, 3D CRFs were also used for resolution refinement.

Although all these CNN methods have achieved promising results, they suffer from the limitation of insufficiently learning both local and global contextual information between pixels. Therefore, models such as the CRF are implemented to embed the spatial contiguity in the output maps. Xue [20] proposed a multi-scale L_1 loss function to force the network to learn both global and local features, for capturing long- and short-range spatial relationships between pixels. However, it requires more computational cost and memory for the network, when the image size gets too large. Additionally, there is the problem of label and class imbalance, which deteriorates the segmentation results. To mitigate these problems, we propose the coarse-to-fine adversarial network for ENKL segmentation, which achieves high segmentation accuracy by locating lesion zones, while cropping unnecessary information reduces computational cost. The multi-scale L_1 loss function is employed to enforce the learning of hierarchical features in a more straightforward and efficient 2D network manner. We make use of a coarse-to fine approach in our framework. In the coarse stage, we first train a shallow U-Net for reducing the image size of our dataset by roughly localizing the target lesion from the whole PET and CT dataset. Then, a modified adversarial network is trained from the sub-volumes sampled from the ground truth bounding boxes of the target lymphoma. We refer to this as coarse-to-fine framework, which is designed to achieve better segmentation performance and relieve memory issues as well as class imbalance problems. The coarse step removes a large amount of the unrelated background region. Then, due to the reduced region size, we simplify the task for the fine step, where the network learns patterns to distinguish the lymphoma from the background. In specifically, the network exploits local context in order to obtain more accurate segmentation results.

Standard deep learning based segmentation models usually produce probability estimates, if a pixel belongs to a certain segmentation label. This kind of approach typically lacks of uncertainty quantification [21], [22]. In medical applications, it can lead to false conclusions when the uncertainty estimates are not well quantified and calibrated. Because of the mathematical complexity of Bayesian approaches in traditional deep learning, Gal and Ghahramani [23] proposed a simple approach of uncertainty estimation. There, they train a dropout network [24] and extract MC samples from the prediction by using dropout at test time. This approach produces an approximation of the posterior of the network's weights. Tanya et.al [25] developed a CNN for Multiple Sclerosis (MS) lesion segmentation and provided an augmentation for providing four different voxel-based uncertainty measures based on MC dropout. This theory has been applied to many medical applications [26]–[28], but there is limited research on zone-based normalization uncertainty analysis to judge the superiority of an algorithm. This paper proposes quantitatively multi-zone uncertainty criteria (lesion-based and background-based) to clearly understand the main source of uncertainty in the background and the lesion zones.

The main contributions of the paper are the following:

1. To the best of our knowledge, this paper is one of the first deep learning studies on computer-aided diagnosis systems for an ENKL dataset. We propose a coarse-to-fine adversarial network for ENKL semantic segmentation. The network architecture is specifically designed to address difficulties of the ENKL dataset. The coarse stage acts as a dimensionality reduction to roughly locate the lesion bounding box and crops redundant information to facilitate the fine segmentation, which is crucial to reduce segmentation time and avoid memory problems. Experiments show that the coarse-to-fine mechanism is effective to improve the segmentation accuracy by reducing the background information.

2. The fine segmentation is an end-to-end adversarial network with a generator and a discriminator part. Spatial context information and hierarchical features are exploited by introducing a multi-scale L_1 loss function in both the generator and discriminator parts without further smoothing of predicted label maps using CRFs. In order to verify the effectiveness and superiority, the presented method is compared with the start-of-the-art U-Net. Extensive experiments demonstrate that the proposed method achieves comparable or better results in terms of volume similarity, lesion overlap and spatial distance than the state-of-the-art CNN-based U-Net architecture.

3. We present an exploration of zone-based uncertainty estimates based on Monte Carlo (MC) dropout in the context of deep networks for medical image segmentation. We proposed uncertainty criteria including two parts: lesion-based and background-based uncertainty criterion. The pixel-based deviation map computed by the MC dropout is normalized to the specific interval using the linear normalization. The lesion-zone pixel-wise sum of normalization map is calculated as the definition of lesion-based uncertainty criterion while the background-zone pixel-wise sum is defined as the background-based uncertainty criterion. A clear understanding of the main source of uncertainty in the respective zones is crucial for a quantitative evaluation of an algorithm's stability. Furthermore, it makes it possible to permit subsequent optimization by engineers and revision by clinicians.

In the following, we present a brief overview of the coarse-to-fine adversarial network architecture and uncertainty theory. We then propose and evaluate our model architectures for our ENKL dataset. Finally, we compare three different network frameworks and show that the proposed network performs much better than the state-of-the-art U-Net segmentation method in terms of volume similarity, lesion overlap, spatial distance and lesion-based stability.

II. Methodology

A. ENKL dataset and preprocessing

Extranodal natural killer (NK)/T-cell lymphoma, nasal type (ENKL) is a predominantly extranodal lymphoma associated with EpsteinBarr virus (EBV) [29], [30]. ENKL is much more common in Asia and Latin America than in the USA and Europe [31], [32]. From June 2011 to March 2018, 83 patients with recently diagnosed ENKL underwent a 18F-FDG PET/CT scan for initial staging at Shanghai Ruijin Hospital.

The diagnosis is according to the World Health Organization pathologic classification. The diagnosis was established by histopathological examination of biopsy tissue from nodal or extranodal disease sites. All patients were followed up at least 12 months. All procedures performed in the study were in accordance with the ethical standards of the committee from Ruijin Hospital, Shanghai Jiaotong University, School of Medicine. 18F-FDG PET/CT was performed on a Discovery STE16 system (GE Healthcare, Waukesha, Wisconsin, USA). Patients were required to fast for at least 6 h before undergoing imaging, and the serum glucose level was kept under 7.0 mmol/L. A image was obtained about 1h after intravenous administration of 5-6 MBq of 18F-FDG per kilogram of body weight.

The difficulties in the dataset mainly stem from three aspects: 1) The large variations in the shape, size and location of the lymphoma. 2) Due to the large image sizes, network suffers from memory size, complicating image processing approaches that take the whole volume as input. 3) The images coming from PET and CT are not identical in their sizes. Consequently, this complicates the straightforward usage of information from two modalities for boosting segmentation accuracy. Details on the properties of the dataset are listed in Table I. Therefore, we perform a rigid co-registration using the P-Mod 3.9 to get identical image sizes. The Normalized Mutual Information was chosen as the dissimilarity function because of the good performance in many multi-modality situations. The interpolation method was the trilinear method. The value of the function tolerance was set as 1.0E-5. This process is shown in Fig. 1. The dataset consists of 83 subjects in total, including both PET and CT images, along with the corresponding binary masks, annotated by three experts who have more than 8-year experience. Two radiologists outlined the shape of lesions based on two modalities information. The other radiologist further checked and modified the original ground-truth maps to decrease the intra-raters errors.

TABLE I
THE DATASET CONSISTS OF 83 SUBJECTS IN TOTAL, INCLUDING BOTH PET AND CT IMAGES, ALONG WITH THE CORRESPONDING BINARY MASKS, ANNOTATED BY EXPERTS BASED ON THE PET AND CT IMAGES AFTER THE CO-REGISTRATION PROCESS.

ENKL lymphoma data	Patients	Size
CT	83	$47 \times 512 \times 512$
PET	83	$47 \times 128 \times 128$
Masks by experts after co-registration	83	$47 \times 512 \times 512$

Image preprocessing plays a crucial role in the deep learning framework. Firstly, for minimizing the variation of voxel intensities, Gaussian normalization is implemented for each 3D scan. Secondly, we augment the training set to achieve invariance and robustness. While normalization is applied for training and testing data, including both PET and CT, data augmentation is only used during training.

B. Data Augmentation

Data augmentation is an effective way to equip deep networks with invariance and robustness properties, when training data is limited. Biomedical images from different subjects and scanners show variations in head orientation, voxel size and lesion distribution. Therefore, rotation and scale invariance as well as robustness to shear transformation is of capital importance here. For each axial slice, three transformations including rotation, shear mapping and scaling are applied, each within a specific parameter range. The parameter range represents the variation in different aspects between subjects in clinical practice. For instance, rotation of the brain is in the range of $[-15°,15°]$. Table II lists the parameter range for each of the three transformations. It should be noted that the scaling used for training is in the range of [0.9, 1.1], representing the range of voxel size ratios in the training dataset. This indicates the robustness of our approach, but also leaves room for improvement in future studies exploring the optimal data scaling during training. After data augmentation, we obtain a dataset four times larger than the original one.

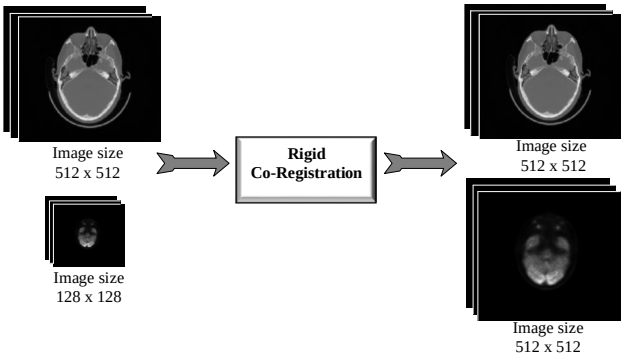


Fig. 1. Illustration of the rigid co-registration for achieving identical image sizes in PET and CT.

TABLE II
PARAMETERS RANGES FOR DATA AUGMENTATION

Data Augmentation	Rotation	Shearing	Scaling(x,y)
Parameters	$[-15°,15°]$	$[-18°,18°]$	[0.9,1.1]

C. Coarse-to-fine network architecture

In this section, we introduce our coarse-to-fine adversarial segmentation network, consisting of a coarse and a fine stage. The coarse stage is a 4-layers shallow U-Net to quickly get a coarse bounding box around the lesion. As shown in Fig. 2, both PET and CT are fed into the coarse stage as two-channel input. The coarse part consists of a contracting path to capture context and a symmetric expanding path, which enables precise localization. Skip connections between the contracting path and the expanding path are employed. Two convolutional layers are repeatedly applied, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. For the first two convolutional layers, we change the kernel size from 3×3 to 5×5, which claims that a larger kernel size enhances a network's capability to handle different transformations. After the preliminary coarse segmentation, we detect the lesion and extract a volume of 128×128 around it as shown in Fig. 3.

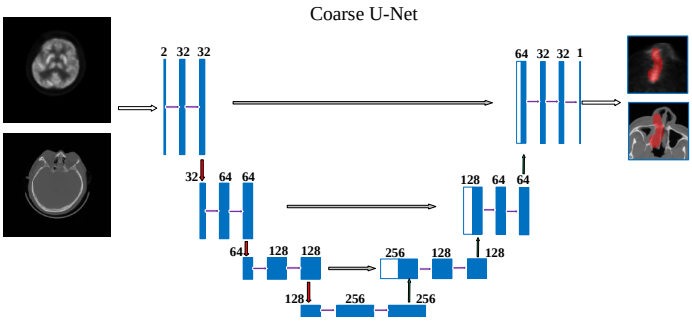


Fig. 2. The coarse net: 2D Convolutional network architecture to locate the lesion.Different operations are denoted by different arrows.The multi-channel feature maps are shown in blue and the copied feature maps are shown in gray. The digit above the feature maps denotes the number of channels.

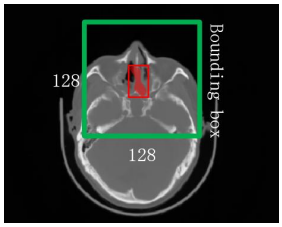


Fig. 3. The bounding box (128 × 128) defined around the lesion location after the coarse segmentation stage. The bounding box in green having the lesion rectangle in red as its center.

The fine network is an end-to-end adversarial network architecture with a multi-scale L_1 loss function, as shown in Fig. 4. It consists of two parts: the generator network G and the discriminator network D. The task for the generator network G is to generate binary masks for the input dataset. The task for the discriminator network D is to distinguish two different input types: the ground truth and the prediction label map. During the adversarial learning, the generator network is forced to learn more accurate prediction maps. The G and D networks are alternately trained by backpropagation in an adversarial fashion. G aims to minimize the multi-scale L_1 loss, while D maximizes this loss function. A fully convolutional encoder-decoder structure similar to the U-Net is used for the generator network G. In the decoder step, for the convolutional kernel, we choose relatively large kernel sizes of 11, 9, 7, in order to achieve a large reception field. Residual blocks, dropout layers and batch normalization are added to the network for preventing overfitting and incorporating of a spatial prior. The discriminator network D has a similar structure as the decoder in G, but in a reverse direction and without residual blocks. The loss function used for the coarse network as well as for the comparison network is based on the dice score and is defined as [33]:

$$minL_{Dice} = \frac{2\sum_{n=1}^{N} p_i g_i + s}{\sum_{n=1}^{N} p_i^2 + \sum_{n=1}^{N} g_i^2 + s} \tag{1}$$

where g_i and p_i represent the ground-truth and predicted probabilistic pixel, respectively. The rest term s ensures stability by avoiding the division by 0. We set s to 1 in our experiments, where the entries of g and p are all zeros.

Conventional GANs [34] typically use loss functions in an adversarial manner:

$$\begin{aligned} &\min_{\theta_G}\max_{\theta_D}\xi\left(\theta_G,\theta_G\right) \\ &= \mathbf{E}_{\mathbf{x}\sim\mathbf{P_{data}}}[\log \mathbf{D}(\mathbf{x})] + \mathbf{E}_{\mathbf{x}\sim\mathbf{P_z}}\log(\mathbf{1}-\mathbf{D}(\mathbf{G}(\mathbf{z}))) \end{aligned} \tag{2}$$

where x is a real image from the unknown data distribution P_{data}, and z is a random input for the generator, following a probability distribution. θ_G and θ_D represent the parameters for the generator and the discriminator in a GAN.

Given a dataset with N training images x_n and the corresponding ground truth maps y_n, the multi-scale loss function is expressed as:

$$\begin{aligned} &\min_{\theta_G}\max_{\theta_D}\xi\left(\theta_G,\theta_D\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\ell_{mae} f_D(x_n \bullet G(x_n)), f_D(x_n \bullet y_n)) \end{aligned} \tag{3}$$

Where ξ_{mae} is the mean absolute error(MAE) and $x_n \bullet G(x_n)$is the entry-wise product of the original images and the segmentation prediction, while $x_n \bullet y_n$ is the pixel-wise multiplication of the original images and the ground truth. The mean absolute error is defined as:

$$\ell_{mae}(f_D(x), f_D(x^{'})) = \frac{1}{L}\sum_{i=1}^{L}\left\| f_D^i(x) - f_D^i(x^{'})\right\|_1 \tag{4}$$

where L is the number of layers in the discriminator network and $f_D^i(x)$is the feature map at the ith layer of the discriminator network.

D. Uncertainty theory

Dropout is a way of Bayesian approximation. During training, the input channels x and the corresponding ground truth lesion labels Y are used to learn the weights θ of the network. To capture the uncertainty character in the model, a prior distribution is placed over θ and an estimate of the posterior $p(\theta|X,Y)$ is calculated. An analytical computation of this prior is intractable, but variational methods can approximate it with a parameterized distribution $q(\theta)$ by minimizing the Kullback-Leibler (KL) divergence [35]:

$$q^*(\theta) = argminKL(q(\theta)||p((\theta|X,Y))_{q(\theta)} \tag{5}$$

According to [23], Yarin et al. declare that minimizing the cross-entropy loss of a network with dropout applied after each layer of weights is equivalent to the minimization of the KL-divergence. In order to analyze the reliable capability of the network, the Monte Carlo dropout method is introduced here. Additionally, a novel corresponding uncertainty evaluation criterion is proposed to measure the networks resistance to the epistemic uncertainty according to the variance map, which is directly obtained from the probability map. The variance map intuitively reflects the prediction fluctuation of the network architecture. The uncertainty sources are mainly from two aspects, background and lesion. The uncertainty from the lesion-zone causes the lesion prediction error and severely affects many important criteria such as the Dice Score and Sensitivity whose evaluation policy relies on the lesion region

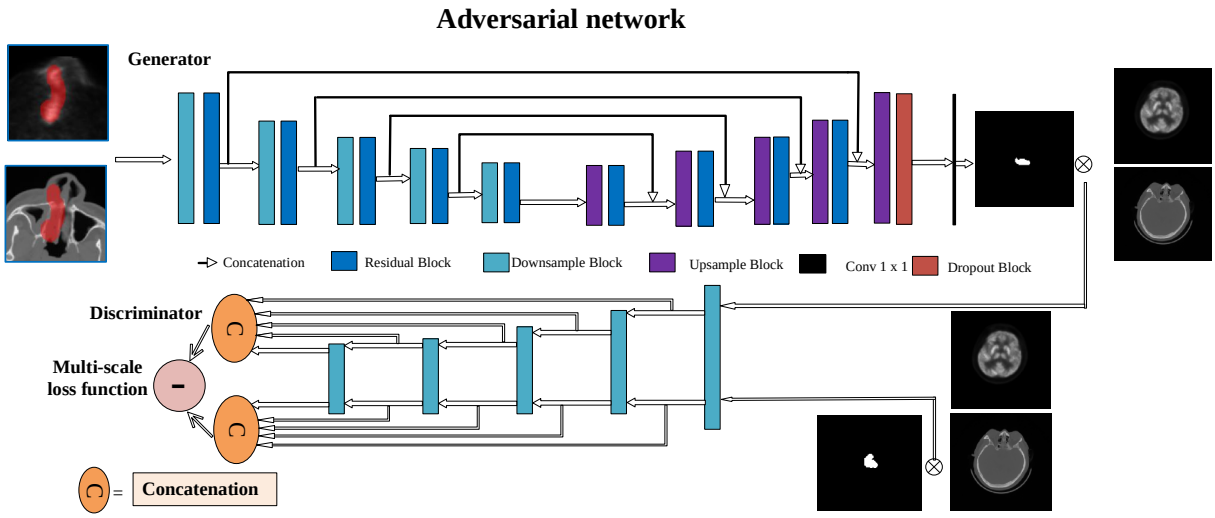


Fig. 4. Architecture of the fine network: Adversarial network combining a generator and a discriminator network. Masked images are calculated by a pixel-wise multiplication of a label map and two channels of an input image.

and its boundary pixels. In sharp contrast to the uncertainty from the lesions, the uncertainty from the background mostly affects the criteria, which rely on pixels predicted as non-lesion pixels, such as specificity (true negative rate), which measures the proportion of actual negatives that are correctly identified. The lesion-based criteria (such as average volume difference) are insensitive to the uncertainty from the background, especially for a low ratio of number of lesion pixels to number of background pixels. Consequently, the paper proposes a novel principle to evaluate the lesion-based uncertainty by calculating the variance map lying in the lesion and its boundary region, and converting the value to the same magnitude. The lesion-based or background-based uncertainty is defined as:

$$U_{zone} = \sum_{i \subset S}^{m} \frac{\sum_{j}^{MC} (X_j - \overline{X})^2}{MC} \tag{6}$$

where i and m are the pixels from the set S lesion or the background of the ground truth, respectively. MC represents the iteration number of the Monte Carlo dropout method, $\overline{X}$ represents the mean map calculated by the prediction probability map and x_j is the j*th* probability map of the MC dropout method. Afterwards, linear normalization is implemented as a post-processing step for the uncertainty map, which is achieved from Eq. (6):

$$E_{zone} = \frac{O_{max} - O_{min}}{U_{max} - U_{min}} \bullet (U_{lesion} - U_{min}) + O_{min} \tag{7}$$

where O_{max}, O_{min}, U_{min} and U_{max} are the maximum and minimum of the specific normalization interval and the original uncertainty map. After defining the uncertainty evaluation criteria, the lesion-based and background-based uncertainty analysis is conducted in the following session.

III. Experiments and Results

The whole dataset is divided into five folds, and a K-fold-cross-validation (K=5) is implemented to comprehensively evaluate the performance of the proposed network architecture. The number of patients in each fold is 16, 16, 17, 17, 17, respectively. For each model, we used approximately 80% of the dataset for training and the remaining 20% for testing and the training set was further split into a training and validation subset in a ratio of 4:1 (where the validation set was used for parameter tuning purpose).

In this section, for proving the remarkable performance of our method, we compare the performance of the proposed fine network with the performance of a U-Net based approach for white matter hyperintensities segmentation, which won the WMH Segmentation Challenge in 2017 [36]. We present the experiments performed with three methods: our proposed method, a U-Net using a bounding box 128x128 (U-Net 128) [36] but without adversarial network, a U-Net using the original images of size 512x512 (U-Net 512) and 3D-UNet [37]. For each method, a 5-fold cross validation was applied, where each fold is used as a test case.

A. Evaluation Metrics

Given a ground-truth segmentation map G and a prediction map R generated by the algorithm, four evaluation metrics are introduced to evaluate the algorithm performance. The metric dice similarity coefficient (DSC), which measures the overlap between the ground truth and the prediction map, is defined as:

$$DSC = \frac{2\,|\mathbf{R} \cap \mathbf{G}|}{(|\mathbf{R}| + |\mathbf{G}|)} \tag{8}$$

Sensitivity (also called the true positive rate) measures the proportion of actual positives that are correctly identified. Assume P is the number of all pixels that are real lesion pixels (real positive) and N is the number of all pixels that are real non-lesion pixels (real negative). P consists of the true positive pixels TP, that are the pixels correctly predicted as lesion pixels, and the false negative pixels FN, that are

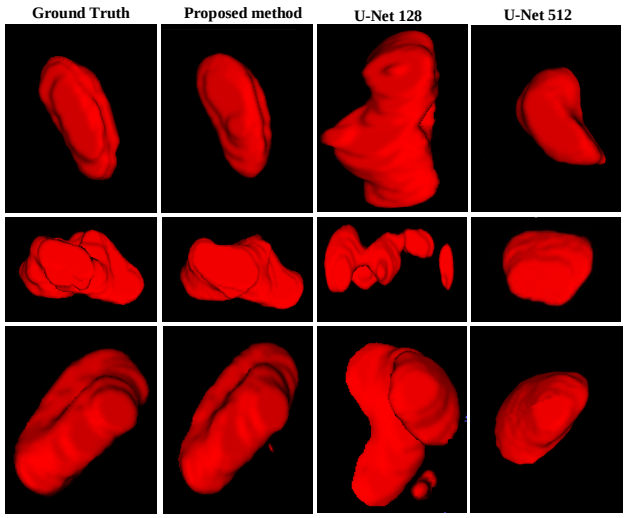


Fig. 5. 3D view of sample lesion segmentations obtained by the proposed method, the U-net 128 and the U-net 512(right) as well as the corresponding ground truth (left).

the pixels wrongly predicted as non-lesion pixels. Then, the sensitivity is defined as:

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{9}$$

For the purpose of measuring the spatial similarity between the prediction map and the ground truth, the Hausdorff distance is introduced here as a key criterion. The Hausdorff distance (95*th* percentile) is defined as:

$$\begin{aligned} H(G,P) =& max(supinf_{x\in G, y\in P} d(x,y), \\ & supinf_{y\in P, x\in G} d(x,y)) \end{aligned} \tag{10}$$

where $d(x, y)$ denotes the distance of x and y, sup denotes the $supremum$ and inf denotes the $infimum$. This measures how far two subsets of a metric space are from each other. For this comparison, it is modified to obtain a robust version by using the 95*th* percentile instead of the maximum (100 percentile) distance. As a result, a closer distance (smaller value) of the spatial geometry represents a higher spatial similarity and therefore, a better performance.

The average volume difference is also a typical criterion to evaluate the volume difference (such as over-segmentation or under-segmentation) between ground truth and prediction map. Let V_G and V_P be the volume of lesion regions in G and P, respectively. Then the Average Volume Difference (AVD) is defined as:

$$AVD = \frac{|V_G - V_P|}{V_G} \tag{11}$$

A smaller difference (smaller value) in the volume size of the ground truth and the prediction map represents a better performance and a higher volume similarity.

B. Comparison with State-of-the-art

The four metrics from the 5-fold cross validation experiments are listed in Table III. It indicates that the 3D-UNet performs a little better than the U-Net 128, but inferior than the proposed method. Fig. 5 shows sample results of the lesion segmentation. Obviously, the proposed method achieves superior results to the U-Net 128 in terms of lesion volume and shape, when comparing the respective segmentations to the ground truth. Moreover, the U-Net 128 tends to over-segment the lesions, which worsens the overall performance. Four different metrics are used here to compare and rank the methods. Those metrics evaluate the segmentation performance in different aspects, such as the volume similarity, the lesion overlap and the spatial distance. Besides, Fig. 6 shows the distributions of the segmentation performances on the test patients of the 5-fold experiments using three different methods.

We claim that the improvement over the traditional approaches is contributed by the coarse-to-fine setting and the adversarial network. The coarse-to-fine stage aims to crop the unnecessary information and roughly locate the lesion region to relieve the class unbalance problem. Interpreting the Dice scores as percentages, the mean Dice score obtained with the U-Net 128 is almost 5% higher than the one achieved by the U-Net 512. This shows that the coarse-to-fine architecture is effective and beneficial in assisting to predict a more accurate segmentation after extracting the bounding box around the lesion. The presented method achieves a more accurate segmentation result in terms of the dice score metric compared with the U-net 128. This shows that the adversarial network further improve the accuracy of the current state-of-the-art U-Net architecture. From the comparisons of three other metrics between the two U-Net-based methods, it shows that the coarse-to-fine framework largely improves the Hausdorff Distance and Dice Similarity Coefficients of segmentations. Furthermore, the adversarial fine network makes a great contribution to improve the average volume similarity when compared with the fine U-Net 128.

Fig. 6 shows that our proposed network generates less outliers than the other methods calculated by U-Net-based methods. This demonstrates that our proposed method can provide more stable and reliable segmentation results than the other two U-Net frameworks.

C. Modality Analysis

Different modalities contain different information, which is crucial to accurately predict the lesion region. For the Extranodal natural killer (NK)/T-cell lymphoma (nasal type), PET and CT images were collected for the lesion segmentation.Comparison experiments of different input modality for the system are performed. Specifically, the models are trained on the data from single modalities (either PET or CT) and the combination of two modalities. The results calculated by the proposed method from the four metrics are listed in Table IV. It shows that the two modalities as input, which provide more latent information, can obtain much better segmentation results on four metrics. Comparing the results from the single modalities, using only PET images achieves more accurate segmentation results than using only CT images as input. This indicates that PET modality plays the key role in the ENKL segmentation task, while the CT modality just provides complementary information to boost the segmentation

TABLE III
COMPARISON OF THE FOUR METRICS(DICE SIMILARITY COEFFICIENT,SENSITIVITY,HAUSDORFF DISTANCE,AVERAGE VOLUME DIFFERENCE) FROM THE THREE METHODS,↓ INDICATES THAT THE SMALLER VALUE REPRESENTS BETTER PERFORMANCE.

Metrics	Dice Similarity Coefficients	Sensitivity	Hausdorff Distance ↓	Average Volume Difference ↓
Proposed method	**0.7115±0.132**	**0.7472±0.185**	**5.9781**±9.317	**0.3711±0.421**
3D-UNet	0.6944±0.136	0.7446±0.186	6.8202±10.027	0.4566±0.463
U-Net 128	0.6798±0.177	0.7332±0.229	6.6061±**8.199**	0.6018±0.706
U-Net 512	0.6282±0.214	0.7006±0.286	21.0280±27.451	0.5594±0.495

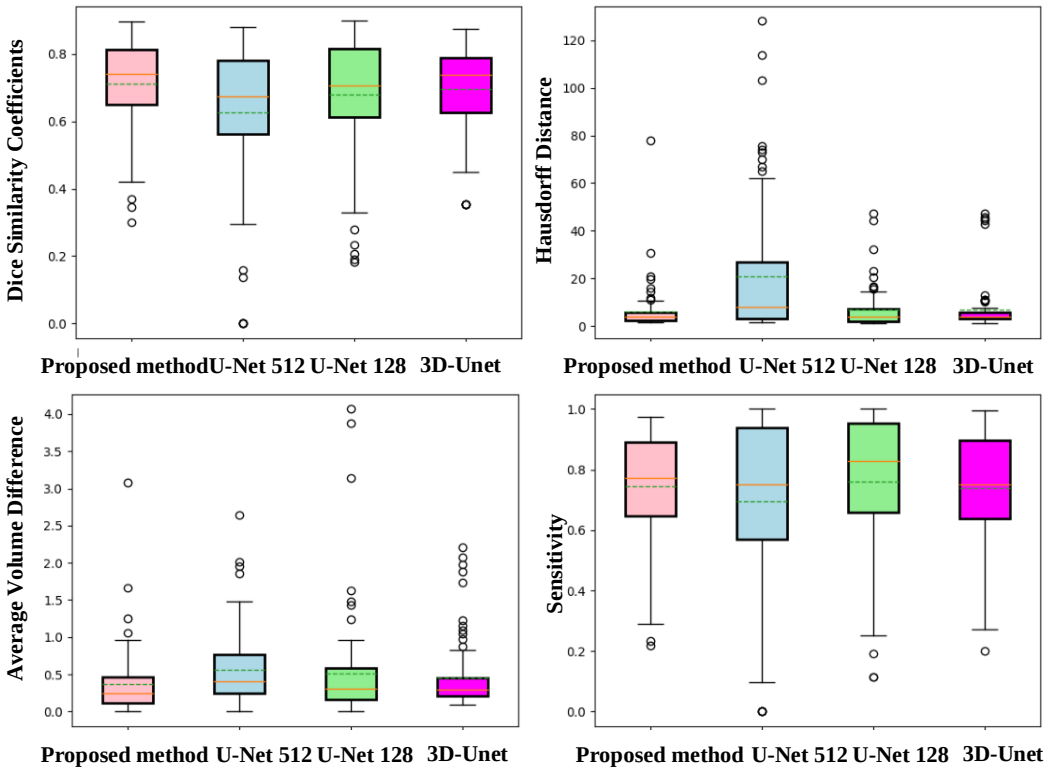


Fig. 6. Boxplot showing the four evaluation metrics of the three methods for the 5 fold test patients. The red line represents the mean value, the green dashed line represents the median value and the circles represent outliers.

accuracy. Fig. 7 shows the distributions and outliers of the four metrics obtained by three modality schemes for the 5 fold test patients. We can see that combining two modalities produces less outliers and therefore, provides more stable and more accurate segmentations, compared to the single modality.

D. Uncertainty evaluation

Uncertainty analysis is a crucial criterion to evaluate the stability of neural network architectures and to improve the understanding and quality of computer-assisted methods used in medical applications. Additionally, we show that the uncertainty criteria model can be combined with standard Monte Carlo dropout Bayesian neural networks to characterize the uncertainty of model parameters. The dropout layer is equally embedded into the last layer of the proposed adversarial network and the U-Net 128. The uncertainty level of the dropout layer is set to 0.2. The main source of the uncertainty is divided into two parts (the lesion-based and the background-based uncertainty), where the uncertainty level calculated by different methods is different, even for the same dataset. It is vital to provide a reliable and quantitative uncertainty evaluation where the uncertainty exactly comes from. Furthermore, the quantitative uncertainty estimates for the predictions permit subsequent revision by clinicians. The lesion-based uncertainty is more critical compared to the background-based uncertainty because it is directly linked to the precision of lesion volume and location. A higher level of lesion-based uncertainty is more likely to make a wrong prediction on the lesion volume. Therefore, the lesion-based uncertainty is reasonable to be regarded as an important performance evaluation criterion of the segmentation models. In this section, we compare the uncertainty criterion of the proposed method with the fine U-Net 128, in order to quantitatively analyze the main uncertainty source for background and lesions.

Fig. 8 shows the uncertainty probability maps of the proposed method and the U-Net 128 with the corresponding ground truth. It is obvious that the main uncertainty sources of the two methods are fundamentally different. There, the

TABLE IV
COMPARISON OF THE FOUR METRICS(DICE SIMILARITY COEFFICIENTS,SENSITIVITY,HAUSDORFF DISTANCE,AVERAGE VOLUME DIFFERENCE) CALCULATED BY THE PROPOSED METHOD FROM THE THREE MODALITY SCHEMES. ↓ INDICATES THAT THE SMALLER VALUE REPRESENTS BETTER PERFORMANCE

Metrics	Dice Similarity Coefficients	Sensitivity	Hausdorff Distance ↓	Average Volume Difference ↓
CT	0.4811	0.5718	9.6206	1.3926
PET	0.6857	0.7066	7.1141	0.8805
CT+PET	**0.7115**	**0.7472**	**5.9781**	**0.3711**

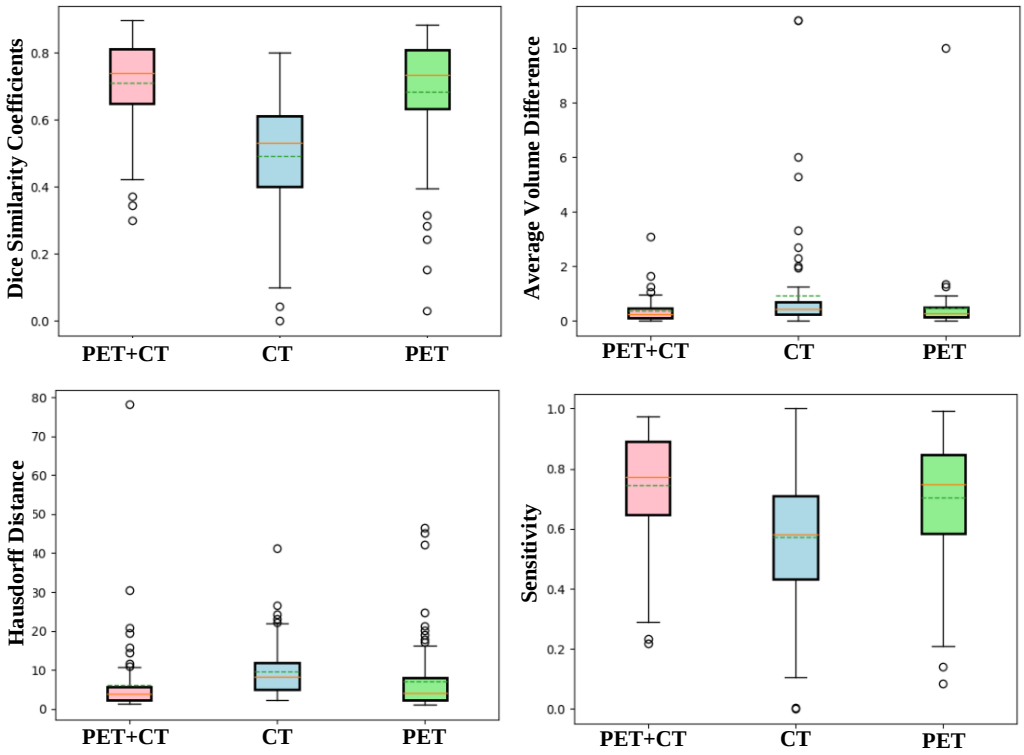


Fig. 7. Boxplot showing the four evaluation metrics of the different modality schemes for the 5 fold test patients. The red line represents the mean value, the green dashed line represents the median value and the circles represent outliers.

darkest zones represent the lowest uncertainty level and the brightest zones the highest uncertainty level. The main uncertainty source of the proposed method comes from the background, while the uncertainty of the U-Net 128 mainly comes from the lesion zones and its boundaries. In order to achieve a quantitative analysis of the uncertainty, the uncertainty is normalized to [0, 255]. The mean pixel number, the uncertainty score and its corresponding ratio are listed in Table V and VI for lesion and background zone, respectively. Regarding the lesion-based uncertainty, the uncertainty score of the proposed method (0.064) is much smaller than the one of the U-Net 128 (25.572). When it comes to lesion-stability, the proposed method is much better than the U-Net 128. Consequently, the predictions of the lesions are more stable, which improves the performance for lesion overlap and volume similarity. Having a look at the background-based uncertainty scores, we can see that the score of the U-Net 128 is almost one third of the score of the proposed method. Therefore, the U-Net performs slightly better than the proposed method in terms of background stability. The main uncertainty source of the proposed method results comes from the background (0.977). Therefore, we state another main contribution, since the adversarial fine network performs more stable for the lesion-based prediction, even though the U-net 128 performs slightly better for the background uncertainty.

IV. DISCUSSION

In this work, we propose a coarse-to-fine adversarial network, which is specifically adapted to the segmentation task and produces superior accuracy in terms of volume similarity, lesion overlap and spatial distance. With our coarse-to-fine mechanism incorporating an adversarial network to locate the lesion bounding box, we not only avoid memory problems and save computation time, but also largely improve the segmentation performance. We addressed the problem of unstable pre-training in the adversarial network for segmentation [38] by introducing the multi-scale feature loss. This measures the difference between the segmentation prediction and the ground

TABLE V
QUANTITATIVE UNCERTAINTY ANALYSIS OF THE LESION-BASED CRITERION. SMALLER VALUES REPRESENT BETTER AND MORE STABLE PERFORMANCE.

Methods	Mean number of lesion	Mean Uncertainty score	Ratio(Uncertainty score/Lesion)
Proposed method	6059	529	0.064
U-Net	6059	102046	25.572

TABLE VI
QUANTITATIVE UNCERTAINTY ANALYSIS OF THE BACKGROUND-BASED CRITERION. SMALLER VALUES REPRESENT BETTER AND MORE STABLE PERFORMANCE.

Methods	Mean number of background	Mean Uncertainty score	Ratio(Uncertainty score/background)
Proposed method	763989	747003	0.97777
U-Net	763989	222756	0.29157

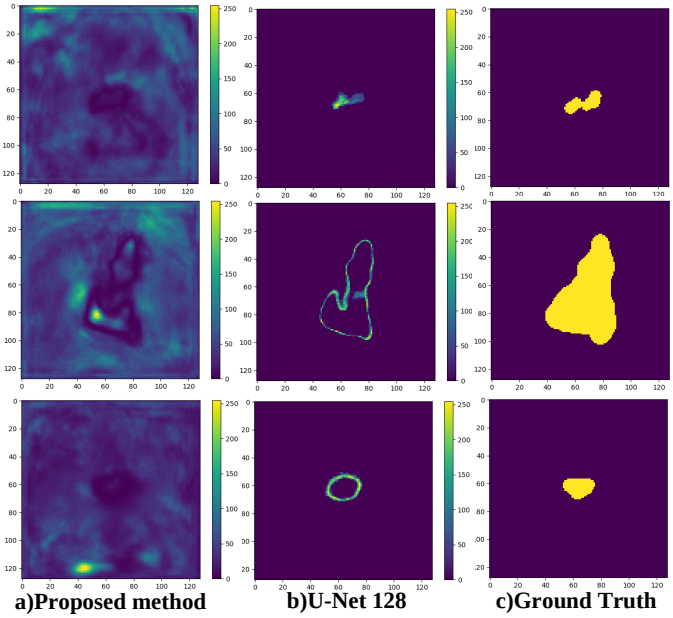


Fig. 8. Qualitative uncertainty analysis for (a) the proposed method and (b) the U-Net 128, while (c) shows the corresponding ground truth. The pixel-wise uncertainty is normalized to the interval [0, 255]. Brighter zones indicate a higher uncertainty character.

truth at multiple layers in the discriminator network, forcing both the generator and the discriminator to learn hierarchical features that capture long- and short-range spatial relationships between pixels. Consequently, this provides sufficient gradients flowing through the discriminator to improve the accuracy and stability of the generator.

In computer vision, modeling uncertainty improves the performance of a standard scene understanding network with no additional parameterization [39]. Moreover, producing only deterministic outputs hinders deep learning from being adopted into clinical routines. Uncertainty estimates for the predictions permit subsequent revision by clinicians. To the best of our knowledge, we present the first exploration of zone-based (lesion-based and background-based) uncertainty estimates based on the Monte Carlo dropout. The uncertainty principles improve the understanding of the algorithm's uncertainty source, coming from the lesion and the background pixels. Additionally, they provide a simple and effective approach to evaluate the stability performance of an algorithm and to exactly verify the uncertainty source. The uncertainty analysis of the proposed adversarial network as well as a U-Net based method show that our method has a much smaller lesion-based uncertainty (0.064) than background-based uncertainty (0.977). However, the uncertainty of the worse performing U-Net mainly comes from its lesion-based uncertainty (25.572), which is much larger than its background-based uncertainty. This induces that a good neural network architecture possesses the property of low level lesion-based uncertainty, which decreases the prediction error stemming from lesion volume and location. This is a crucial criterion to evaluate the superiority of an algorithm. Furthermore, it makes it possible to optimize the uncertainty map and to reduce the uncertainty level by employing post-processing techniques, when clearly understanding the main uncertainty sources from background and lesion zones.

All of the experiments were conducted on a GNU/Linux server running Ubuntu 16.04, with Intel Core i7-6700 CPU and 64GB RAM. The networks were trained on a single NVIDIA Titan-Xp GPU with 12GB RAM.

V. CONCLUSION

In this paper, we presented a coarse-to-fine adversarial network architecture which introduces a multi-scale feature loss to stabilize the adversarial network for biomedical image segmentation tasks. The proposed architecture saves a significant amount of computational memory and time by extracting the lesion bounding box in the coarse stage. Moreover, the results of the four metrics in the experiments (Dice Similarity Coefficient, Sensitivity, Hausdorff Distance and Average Volume Dierence) on the ENKL dataset demonstrate that the proposed method outperforms the state-of-the-art U-net segmentation method. The presented method is a general framework and has the potential to be used in general semantic segmentation tasks.

Additionally, we proposed the zone-based uncertainty criteria (lesion-based and background-based criteria) based on Monte Carlo dropout method. The uncertainty of our deep learning model is quantitatively analyzed, and makes it possible to clearly understand the main uncertainty source. Moreover, the quantitative uncertainty analysis provides clinicians with information permitting them to quickly assess, whether they should accept or reject lesions of high uncertainty.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research. We give our gratitude to the financial support of China Scholarship Council (CSC).

REFERENCES

[1] B. D. Cheson, "Role of functional imaging in the management of lymphoma," *Journal of Clinical Oncology*, vol. 29, pp. 1844–1854, 2011.

[2] W. Chan, W. Y. Au, C. Wong, R. Liang, A. Leung, Y. Kwong, and P. Khong, "Metabolic activity measured by f-18 fdg pet in natural killer-cell lymphoma compared to aggressive b- and t-cell lymphomas," *Clinical Nuclear Medicine*, vol. 35, pp. 571–575, 2010.

[3] P. L. Khong, C. B. Pang, R. Liang, Y. L. Kwong, and W. Y. Au, "Fluorine-18 fluorodeoxyglucose positron emission tomography in mature t-cell and natural killer cell malignancies," *Annals of Hematology*, vol. 87, pp. 613–621, 2008.

[4] X. Hu, H. Li, Y. Zhao, C. Dong, B. H. Menze, and M. Piraud, "Hierarchical multi-class segmentation of glioma images using networks with multi-level activation function," in *International Conference on Medical Image Computing and Computer-Assisted Intervention Workshop BrainLes*. Springer, 2018, pp. 116–127.

[5] J. Cai, Z. Zhang, L. Cui, Y. Zheng, and L. Yang, "Towards cross-modal organ translation and segmentation: A cycle-and shape-consistent generative adversarial network," vol. 52, pp. 174–184, 2019.

[6] Z. Zhao, L. Yang, H. Zheng, I. H. Guldner, S. Zhang, and D. Z. Chen, "Deep learning based instance segmentation in 3d biomedical images using weak annotation." Springer, 2018, pp. 352–360.

[7] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham, and W. Cai, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 991–1004, 2019.

[8] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai, "Multiscale network followed network model for retinal vessel segmentation." Springer, 2018, pp. 119–126.

[9] Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, and B. Menze, "Knowledge-aided convolutional neural network for small organ segmentation," *IEEE journal of biomedical and health informatics*, 2019.

[10] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[11] L. Wang, D. Nie, G. Li, . Puybareau, J. Dolz, Q. Zhang, F. Wang, J. Xia, Z. Wu, J. Chen, K.-H. Thung, T. D. Bui, J. Shin, G. Zeng, G. Zheng, V. S. Fonov, A. Doyle, Y. Xu, P. Moeskops, J. P. Pluim, C. Desrosiers, I. B. Ayed, G. Sanroma, O. M. Benkarim, A. Casamitjana, V. Vilaplana, W. Lin, G. Li, and D. Shen, "Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iseg-2017 challenge," *IEEE Transactions on Medical Imaging*, 2019.

[12] M. E. Celebi, N. Codella, A. Halpern, and D. Shen, "Guest editorial skin lesion image analysis for melanoma detection," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, 2019.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, p. 34313440.

[14] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision*. IEEE, 2015, p. 15201528.

[15] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piece-wise training of deep structured models for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, p. 31943203.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation." Springer, 2018, p. 234241.

[17] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Imageto-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[18] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," vol. 35, pp. 18–31, 2017.

[19] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," vol. 36, pp. 61–78, 2017.

[20] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. l. Huang, "Segan: Adversarial network with multi-scale l1 loss for medical image segmentation," vol. 16, no. 3-4, p. 383392, 2018.

[21] Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, and M. J. Cardoso, "Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions." Springer, 2018.

[22] Z. Shi, G. Zeng, L. Zhang, X. Zhuang, L. Li, G. Yang, and G. Zheng, "Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 569–577.

[23] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*. Springer, 2016, p. 10501059.

[24] S. Nitish, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15(1), pp. 1929–1958, 2014.

[25] T. Nair, D. Precup, L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2018.

[26] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander, "Bayesian image quality transfer with cnns: Exploring uncertainty in dmri super-resolution," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, p. 611.

[27] O. Ozdemir, B. Woodward, and A. A. Berlin, "Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection," in *Conference and Workshop on Neural Information Processing Systems*. Springer, 2017.

[28] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific Reports*, vol. 7, 2017.

[29] E. S. Jaffe, N. L. Harris, H. Stein, and J. W. Vardiman, "World health organization classification of tumours. pathology and genetics of tumours of haematopoietic and lymphoid tissues," *International Agency for Research on Cancer*, 2001.

[30] S. H. Swerdlow, E. Campo, and N. L. Harris, "Who classification of tumours of haematopoietic and lymphoid tissues," *International Agency for Research on Cancer*, 2008.

[31] J. R. Anderson, J. O. Armitage, and D. D. Weisenburger, "Epidemiology of the non-hodgkin's lymphomas: distributions of the major subtypes differ by geographic locations. non-hodgkin's lymphoma classification project," *Annals of Oncology*, vol. 9, p. 71720, 1998.

[32] J. Vose, A. J, and W. D, "International peripheral t-cell and natural killer/t-cell lymphoma study: pathology findings and clinical outcomes," *Journal of Clinical Oncology*, vol. 26, p. 412430, 2008.

[33] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, p. 565571.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*. Springer, 2014, p. 26722680.

[35] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, pp. 859–877, 2017.

[36] H. Li, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, and B. Menze, "Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images," *NeuroImage*, vol. 183, pp. 650–665, 2018.

[37] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2016, pp. 424–432.

[38] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *NIPS Workshop on Adversarial Training*. Springer, 2016.

[39] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference*. Springer, 2017.

5

Weakly supervised deep learning for determining the prognostic value of 18F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type

This chapter has been published as **peer-reviewed journal paper:**.

R. Guo*, **X, Hu***, H. Song*, P. Xu, H. Xu, A. Rominger, X. Lin, B. Menze, B. Li, and K. Shi. "Weakly supervised deep learning for determining the prognostic value of 18 F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type." In: *European journal of nuclear medicine and molecular imaging* (2021), pp. 1–11

Synopsis: This work proposed a weakly supervised deep learning (WSDL) method based on positive–negative unlabeled (PNU) classification to maximize the utility of incomplete and missing follow-up data to improve prognosis prediction of ENKTL. prediction similarity index (PSI) was derived from deep learning features of images and multivariate analysis confirmed PSI to be an independent significant predictor of prediction of progression-free survival (PFS) .

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-021-05232-3

ORIGINAL ARTICLE

Weakly supervised deep learning for determining the prognostic value of ^{18}F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type

Rui Guo[1] · Xiaobin Hu[2] · Haoming Song[2] · Pengpeng Xu[3] · Haoping Xu[4] · Axel Rominger[5] · Xiaozhu Lin[1] · Bjoern Menze[2,6] · Biao Li[1] · Kuangyu Shi[2,5]

Received: 12 August 2020 / Accepted: 1 February 2021

Abstract

Purpose To develop a weakly supervised deep learning (WSDL) method that could utilize incomplete/missing survival data to predict the prognosis of extranodal natural killer/T cell lymphoma, nasal type (ENKTL) based on pretreatment ^{18}F-FDG PET/CT results.

Methods One hundred and sixty-seven patients with ENKTL who underwent pretreatment ^{18}F-FDG PET/CT were retrospectively collected. Eighty-four patients were followed up for at least 2 years (training set = 64, test set = 20). A WSDL method was developed to enable the integration of the remaining 83 patients with incomplete/missing follow-up information in the training set. To test generalization, these data were derived from three types of scanners. Prediction similarity index (PSI) was derived from deep learning features of images. Its discriminative ability was calculated and compared with that of a conventional deep learning (CDL) method. Univariate and multivariate analyses helped explore the significance of PSI and clinical features.

Results PSI achieved area under the curve scores of 0.9858 and 0.9946 (training set) and 0.8750 and 0.7344 (test set) in the prediction of progression-free survival (PFS) with the WSDL and CDL methods, respectively. PSI threshold of 1.0 could significantly differentiate the prognosis. In the test set, WSDL and CDL achieved prediction sensitivity, specificity, and accuracy of 87.50% and 62.50%, 83.33% and 83.33%, and 85.00% and 75.00%, respectively. Multivariate analysis confirmed PSI to be an independent significant predictor of PFS in both the methods.

Conclusion The WSDL-based framework was more effective for extracting ^{18}F-FDG PET/CT features and predicting the prognosis of ENKTL than the CDL method.

Keywords Deep learning · ^{18}F-FDG PET/CT · Extranodal natural killer/T cell lymphoma · Prognosis · Progression-free survival

Rui Guo, Xiaobin Hu and Haoming Song contributed equally to this work.

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence).

✉ Biao Li
lb10363@rjh.com.cn

[1] Department of Nuclear Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2] Department of Informatics, Technical University of Munich, Munich, Germany

[3] State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4] Department of Radiation, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[5] Department of Nuclear Medicine, University of Bern, Bern, Switzerland

[6] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

Published online: 20 February 2021

Introduction

The emergence of artificial intelligence (AI) in the field of medical imaging has led to several breakthroughs [1, 2]. AI has already proven to be advantageous for computer-aided diagnosis in medical imaging, such as for the differential diagnosis of coronavirus disease 2019 [3], skin cancer [4], and diabetic retinopathy [5]. Moreover, it has been developed to help identify imaging-based biomarkers, leading to an improvement in the prognosis of, for example, lung cancer [6, 7], gliomas [8], and nasopharynx cancer [9]. Deep learning is an indispensable part of AI and has been reported to be extremely effective in several medical imaging-related tasks, such as image segmentation, registration, fusion, annotation, computer-aided diagnosis and prognosis analyses, lesion and landmark detection, and microscopic imaging analysis. In such studies, deep learning networks have shown capabilities to automatically extract characteristic features from images, including explicit features, such as the location, distribution, and volume size of lesions, and implicit features at different levels, which were deduced using nonlinear, independent discriminant, and invariant properties. The end-to-end automatic feature extraction does not involve human interaction, and the extracted features are the most implicit. Although the implicit features may be difficult to interpret, they are determinant for the performance of convolutional neural networks (CNNs) and play critical roles in many medical applications [10, 11].

The development of deep learning depends on the availability of a huge amount of data. It is usually challenging to gather a large cohort of patients with survival follow-up after administering the same therapeutic regime. Clinical trials are often associated with incomplete or missing follow-up due to factors such as insufficient follow-up time, patient tolerance, and compliance. This consequently hampers extensive development of deep learning methods for predicting therapeutic prognosis. Maximizing the utility of data gathered by clinical trials is thus a key area of research.

Data augmentation methods such as deformation or generative adversarial networks are often applied to support the development of deep learning methods in the field of image analysis [12]. However, the relationship among imaging, therapy, and survival is more complex than general image analyses. The increased physiological complexity makes it difficult to synthesize meaningful data for training. Furthermore, errors in data preparation may mislead algorithmic development [13]. Weakly supervised classification methods have been established using unlabeled data for regularization under particular distributional assumptions, such as cluster or smoothness assumption; however, the performance relies on the fidelity of the assumption [14–16], and it is usually challenging to find a proper assumption in real application. In contrast, positive–negative unlabeled (PNU) classification [15] is a weakly supervised strategy to deal with a tough task with less knowledge regarding data distribution and, therefore, is less restricted in complex applications. Despite these advantages, because PNU classification is generally applied for classification problems based on low-dimensional feature vectors [15], it is not straightforward to apply this classification to imaging data for survival follow-up in order to improve therapeutic prognosis.

Extranodal natural killer/T cell lymphoma, nasal type (ENKTL) is a rare type of lymphoma with poor survival outcome [17–19]. It constitutes <1% of all lymphomas in Western countries and 3–9% of all malignant lymphomas in Asia [18, 20, 21]. Several investigations have identified that almost all ENKTL lesions are fluorodeoxyglucose (FDG) avid [22, 23]. In patients with ENKTL, the use of ^{18}F-FDG positron emission tomography/computed tomography (PET/CT) for staging is widespread [24–26]. Nevertheless, many contradictions exist pertaining to the value of ^{18}F-FDG PET/CT in predicting the prognosis of ENKTL [22, 27–30]. Some studies [31, 32] have reported that maximum standardized uptake value (SUVmax) of pretreatment ^{18}F-FDG PET/CT is not a statistically significant predictor of overall survival and progression-free survival (PFS). Tumor ^{18}F-FDG uptake cannot reflect the aggressive biologic behavior of ENKTL; however, some studies have reported contradictory results [30, 33]. These studies found that high tumor ^{18}F-FDG uptake was closely associated with unfavorable treatment and survival outcomes. Chang et al. [34] reported that baseline whole-body total lesion glycolysis (TLG) was a good predictor of PFS and overall survival in patients with ENKTL. However, treatment plans were not uniform in these studies, potentially affecting the treatment outcome and predictive value of pretreatment ^{18}F-FDG PET/CT. Prospective research methods have also been used to assess the prognostic value of ^{18}F-FDG PET/CT in ENKTL [31, 35, 36], but considering some uncertainty in the reported results, it remains unclear. A novel solution is accordingly needed. Although deep learning has been advantageous in assisting molecular imaging to optimize therapeutic prognosis [9], it is extremely difficult to develop appropriate deep learning methods for this rare condition with only a limited number of cases.

We herein propose a weakly supervised deep learning (WSDL) method based on PNU classification to maximize the utility of incomplete and missing follow-up data so as to predict the prognosis of ENKTL. We investigated the accuracy and robustness of this data enhancement strategy on a retrospective cohort to test a therapeutic regime for ENKTL.

Material and methods

Patients

One hundred and sixteen-seven patients with histopathologically diagnosed ENKTL from June 2011 to October 2020 recruited at Shanghai Ruijin Hospital were retrospectively collected. Patients who had undergone surgical resection, radiotherapy, chemotherapy, and/or bone marrow transplantation as well as those with other malignancies were excluded. All patients underwent whole-body ^{18}F-FDG PET/CT for initial staging before therapy and were then treated with a therapeutic regime of methotrexate, etoposide, dexamethasone, and pegaspargase (MESA). Eighty-four patients were followed up for at least 2 years. Among them, 49 were sandwiched with radiotherapy for the involved local focus 21 days after two cycles of MESA. They were treated with a linear accelerator producing 6 MV photons. The radiotherapy dose was 50 Gy in 25 fractions, once a day, and 5 fractions every week. Chemotherapy was restarted 28 days after radiotherapy.

Of the 84 patients, 64 were randomly included in the training set; the remaining 20 were unobserved and included in the test set. The ratio of relapse to non-relapse individuals was kept the same in the test and training sets to avoid an extreme imbalance problem. PFS was the major endpoint. Recurrence and lymphoma infiltration were mainly diagnosed based on imaging methods and pathology. The remaining 83 patients without follow-up information or followed up for <2 years were also included in the training set using the proposed WSDL method. To further test the generalization of the WSDL method, data pertaining to the 83 patients were derived from three types of scanners: Scanner 1 (Discovery VCT, GE Healthcare, USA, 39 patients), 2 (Discovery MI, GE Healthcare, USA, 29 patients), and 3 (Biograph Vision, SIEMENS, Germany, 15 patients). The training set thus ultimately comprised 147 patients (Fig. 1).

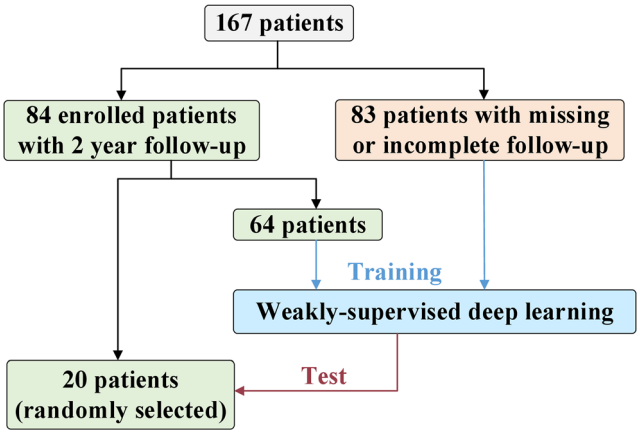


Fig. 1 A flow chart depicting the study plan. ENKTL: extranodal natural killer/T cell lymphoma, nasal type

The clinical features of the 84 patients, including gender, age, serum lactate dehydrogenase levels, Eastern Cooperative Oncology Group (ECOG) score, Ki67, β2-microglobulin, Epstein–Barr virus DNA, and B symptoms, were recorded. Ann Arbor stage, SUVmax, mean SUV (SUVmean), metabolic tumor volume (MTV), and TLG extracted from ^{18}F-FDG PET/CT were also measured. All procedures in the study were performed in accordance with the ethical standards of the committee from Ruijin Hospital, Shanghai Jiao Tong University, School of Medicine. Written informed consent was obtained from all patients before treatment. Among the 84 patients enrolled in the clinical trial, 58 were alive (12 presented with persistent or recurrent disease at the last follow-up), and 26 had died due to a tumor-related disease. The clinical characteristics of patients in the training and test sets have been summarized in Table 1; data pertaining to the 83 patients diagnosed with ENKTL but with missing or incomplete follow-up information are also listed.

^{18}F-FDG PET/CT and preprocessing

Patients were required to fast for at least 6 h before ^{18}F-FDG PET/CT, and the serum glucose level was maintained under 7.0 mmol/L. Whole-body PET from the head to thigh was performed 1 h after intravenously administering 5–6 MBq of ^{18}F-FDG per kilogram of body weight. In case of Scanner 1, PET was performed in the 3D mode with an acquisition time of 2 min per bed position covering the same field as the CT scan. CT was performed using the following parameters: 120–180 mA, 140 kV, gantry rotation speed of 0.8 s, and thick axial section of 3.75 mm. After correcting attenuation (based on CT), scatter, dead time, and random coincidences, PET images were reconstructed using 3D ordered-subset expectation maximization (OSEM) with a Gaussian filter (full width at half maximum of 6 mm), leading to images with voxel size of 5.47 mm. In case of Scanner 2, PET was performed in the 3D mode with an acquisition time of 1.5 min per bed position covering the same field as the CT scan. CT was performed using the following parameters: 120–180 mA, 140 kV, and gantry rotation speed of 0.8 s. PET images were reconstructed using the block-sequential regularized expectation maximization reconstruction algorithm (Q.clear, GE Healthcare, USA), which had a β value of 550 with a 256 × 256 matrix (pixel size = 2.7 × 2.7 mm^2, slice thickness = 2.79 mm). Finally, in case of Scanner 3, CT was performed using the following parameters: 146 mA, 120 kV, and spiral pitch factor of 1. Images were reconstructed using the 3D ordinary Poisson OSEM algorithm, with four iterations and five subsets, application of time-of-flight resolution modeling, and no filtering. The obtained PET images had an image matrix of 440 × 440, pixel size of 1.6 × 1.6 × 1.5 mm, and slice thickness of 2.0 mm. Lymphoma lesions in the training set were manually delineated on the fusion map of PET/CT images using ITK-

Table 1 Clinical characteristics of patients

Characteristics	Training cohort (n=64), no. (%)	Test cohort (n=20), no. (%)	P	Patients with missing or incomplete data (n=83), no. (%)		
				Scanner 1 (n=39)	Scanner 2 (n=29)	Scanner 3 (n=15)
*Gender			0.690			
Male	45 (70.31)	15 (75.00)		28 (71.79)	21 (72.41)	10 (66.67)
Female	19 (29.69)	5 (25.00)		11 (28.21)	8 (27.59)	5 (33.33)
*Age (years)			0.861			
< 60	50 (78.13)	16 (80.00)		24 (61.54)	22 (75.86)	10 (66.67)
≥ 60	14 (21.87)	4 (20.00)		15 (38.46)	7 (24.14)	5 (33.33)
*Primary site of tumor			0.078			
Upper aerodigestive tract	51 (79.69)	12 (60.00)		30 (76.92)	25 (86.21)	13 (86.67)
Non-upper aerodigestive tract	13(20.31)	8 (40.00)		9 (23.08)	4 (13.79)	2 (13.33)
*Ann Arbor stage			0.182			
I–II	51 (79.69)	13 (65.00)		30 (76.92)	23 (79.31)	11 (73.33)
III–IV	13(20.31)	7 (35.00)		9 (23.08)	6 (20.69)	4 (26.67)
*B symptoms			0.213			
Yes	25 (39.06)	9 (45.00)		–	–	–
No	39 (60.94)	11 (55.00)		–	–	–
*ECOG score			0.038			
0	36 (56.25)	6 (30.00)		–	–	–
1	19 (29.69)	7 (35.00)		–	–	–
2–5	9 (14.06)	7 (35.00)		–	–	–
*PINK			0.230			
Low risk (0)	37 (57.81)	11 (55.00)		–	–	–
Intermediate risk (1)	16 (25.00)	1 (5.00)		–	–	–
High risk (2–4)	11 (17.19)	8 (40.00)		–	–	–
**^{18}F-FDG uptake (SUVmax)	13.17±6.90	14.48±4.92	0.432	12.01±6.07	16.40±6.78	21.23±9.06
**Follow-up period (months)	33.70±20.82	38.70±23.96	0.369	–	–	–

*P values were calculated using the chi-squared test for categorical variables and nonparametric test for continuous variables

**Mean ± SD; independent sample t test was used to compare differences in quantitative parameters between the groups

Abbreviations: LDH, lactate dehydrogenase; ECOG, Eastern Cooperative Oncology Group; PINK, prognostic index of natural killer lymphoma; SUVmax: maximum standardized uptake value

SNAP (v3.6.0) by a nuclear medicine physician with 15 years of experience [9].

WSDL for feature extraction

The WSDL method based on Residual Network-18 (ResNet-18) [37] was proposed to predict disease prognosis using a well-exploiting unlabeled dataset (83 patients without follow-up information). The summarized algorithm for the WSDL method is as follows:

Input: 3D volumetric image I of size width × height × depth

Ensure: Image I is a rank 3 tensor

1: Train deep convolutional neural networks (DCNNs) with labeled data to obtain the baseline model
2: Use baseline DCNNs to extract features from labeled and unlabeled data
3: Build the PNU classifier to generate implicit labels for unlabeled data
4: Re-train DCNNs with labeled and unlabeled data to obtain the final prognosis

The ResNet is an artificial neural network that is inspired by the biological neural networks constituting animal brains.

DCNNs were constructed for deep learning feature extraction. They are a simplified version of ResNet-18 and were implemented using the Python Keras package with TensorFlow as the backend. The 83 patients with missing or incomplete follow-up data were included in the training set along with 64 patients with follow-up data. Labels for the 83 patients were implicitly derived using the PNU classifier during the training procedure, leading to maximized prediction probability. Further details are provided in Supplementary Materials.

In total, 128 deep learning features were extracted from the output of the average pooling layer of DCNNs for PET/CT images in the training set, which were grouped into a 16 × 8 feature map for visualization. We herein propose a new biomarker in the form of prediction similarity index (PSI), which is the ratio of the positive predicted probability value to the negative predicted probability value. It was derived from these features to predict the probability of recurrence and non-recurrence. PSI of 1 was used to differentiate between positive and negative predictions. To determine the advantages of the WSDL method, we compared it with the conventional deep learning (CDL) method of our proposed DCNNs trained only on the 64 patients followed up for at least 2 years (Fig. 2).

Statistics

SPSS v23.0 (SPSS Inc., Chicago, IL, USA) and GraphPad Prism 8.0.1 (GraphPad, San Diego, USA) were used for statistical analyses. Univariate analysis using the Kaplan–Meier method was performed for each variable with a potential prognostic value. Time-dependent receiver operating characteristic (ROC) analysis was performed to evaluate the discriminative ability of PSI for the prognostic prediction of ENKTL. PSI-based PFS, prediction sensitivity and specificity, and accuracy of PSI were calculated. Differences in sensitivity and specificity between the WSDL and CDL methods were compared using the Fisher's exact test. The log-rank test was used to compare differences in PFS between the groups (PSI > 1 and PSI < 1). Multivariate analysis using the Cox proportional hazards model was used to assess the independent effects of PSI and clinical parameters of the disease. $P < 0.05$ indicated statistical significance.

Results

Extraction of deep learning features

One hundred and twenty-eight features were extracted from tumor ROIs outlined on ^{18}F-FDG PET/CT scans of each patient using the proposed WSDL method. These ROIs were outlined based on lesion locations and shapes, while non-meaningful background was cut off. The 128 features were grouped into feature maps of 16 × 8 strips. The feature maps of the test set ($n = 20$) have been illustrated in Fig. 3. In general, characteristic differences between relapse and non-relapse patients could be visualized on these maps. The feature maps of the training set ($n = 64$) have been illustrated in Supplementary Figure S1 (relapse) and S2 (non-relapse), whereas those of the 83 patients with incomplete or missing follow-up data and who were imaged using the aforementioned scanners are illustrated in Figure S3. The feature maps of the test set (Figure S4) and training set (Figure S5 for relapse, Figure S6 for non-relapse) with the CDL method have also been illustrated in supplementary figures.

PSI as the prognostic score

Patients with PSI > 1 were considered to show a positive response, while those with PSI < 1 were considered to show a negative response. The ROC curves of the results of the

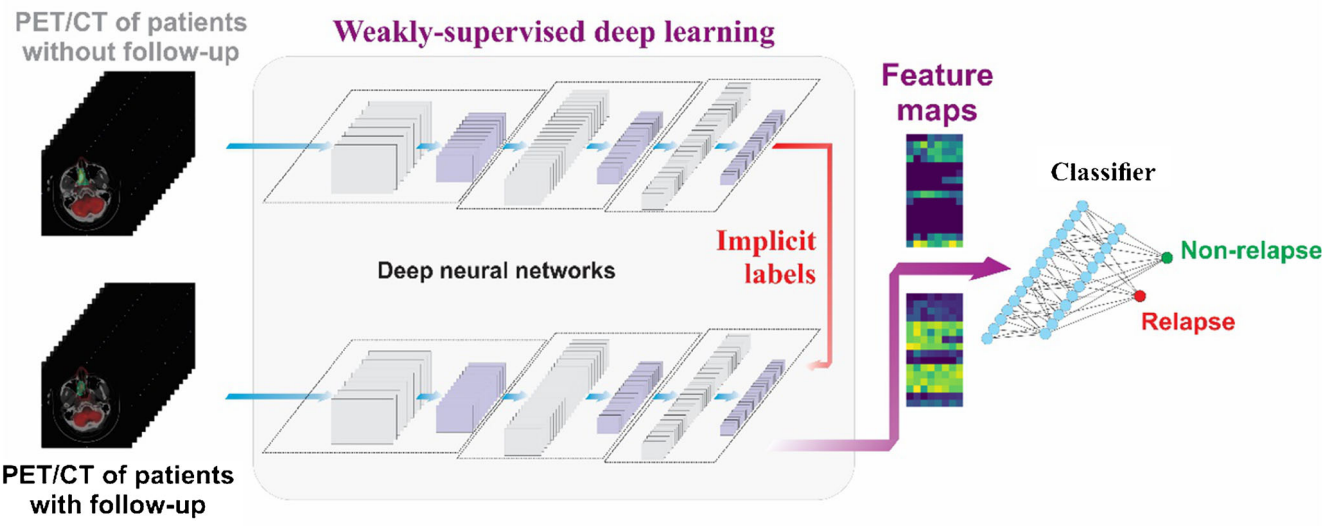


Fig. 2 An illustration of the concept of the proposed weakly supervised deep learning method

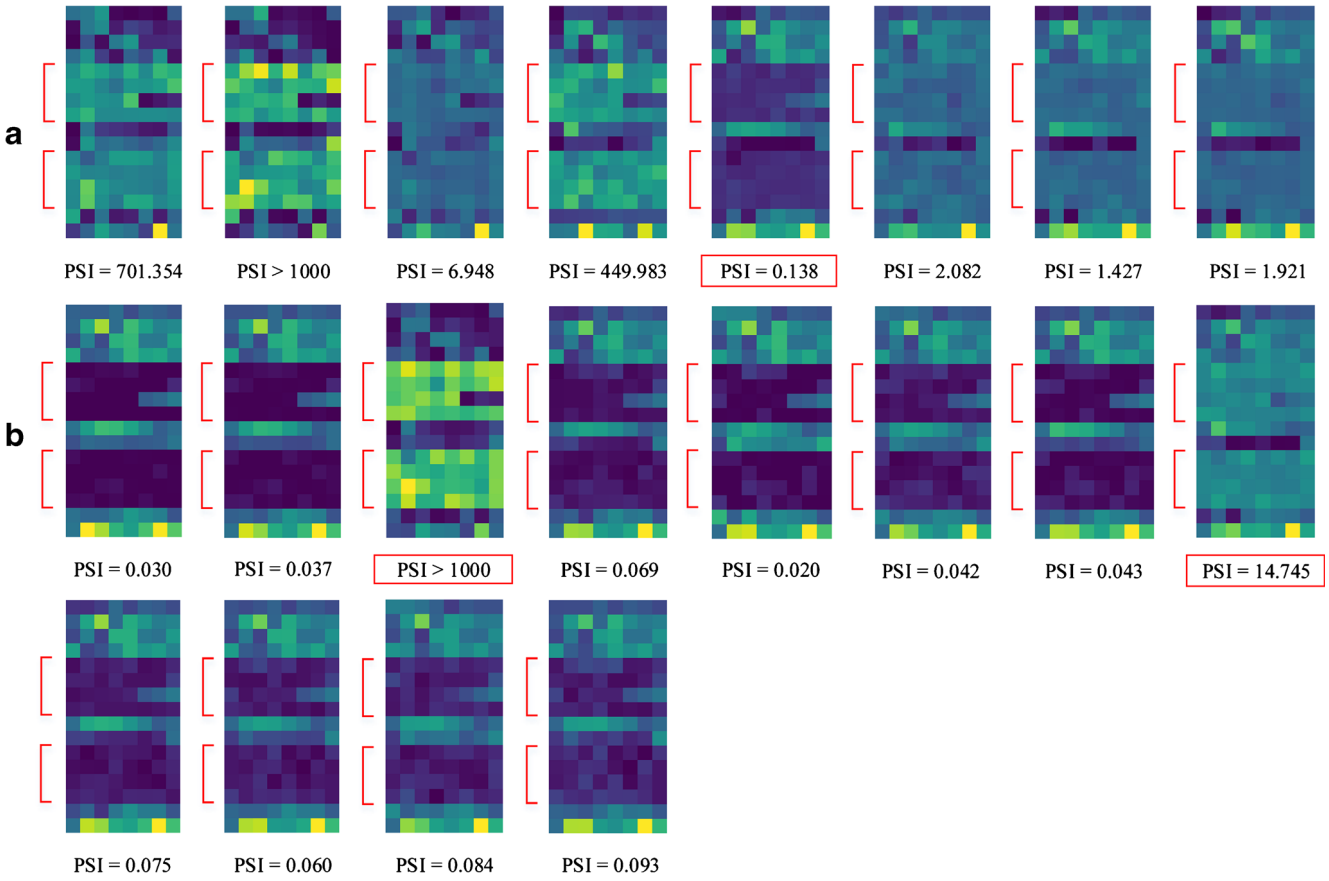


Fig. 3 Visualization of the feature maps (16 × 8) representing 128 features extracted by the proposed WSDL method in the test set. Each strip represents the feature map of a patient. Red arrows indicate the characteristic difference between the (A) relapse and (B) non-relapse groups in the test cohort. PSI results with incorrect predictions have been marked by red boxes

WSDL and CDL methods were compared (Fig. 4). With the WSDL method, in the training and test sets, PSI achieved area under the curve (AUC) scores of 0.986 ($P = 0.000$, 95% CI, 0.957–1.000) and 0.875 ($P = 0.005$, 95% CI, 0.706–1.000), respectively, in the prediction of PFS, while with the CDL method, PSI achieved AUC scores of 0.995 ($P = 0.000$, 95% CI, 0.984–1.000) and 0.734 ($P = 0.083$, 95% CI, 0.479–0.989), respectively (AUC of the training set was calculated only based on data pertaining to the 64 patients). Table 2 shows accuracy and prognosis results. In the training set, the sensitivity of the WSDL method was superior to that of the CDL method (86.7% vs 73.3%, $P = 0.048$), while the methods showed the same specificity (100%). Due to the small number of patients in the test set, a comparison was not feasible.

According to PSI, patients were divided into two groups: PSI > 1 and PSI < 1. The Kaplan–Meier survival analysis method was used to compare differences in PFS between the groups. We observed that patients with low PSI (PSI < 1) showed good prognosis and long PFS, while those with high PSI (PSI > 1) showed poor prognosis and short PFS. Figure 5 shows the Kaplan–Meier curves of PFS according to PSI. The extracted PSI was able to segregate patients in the training set

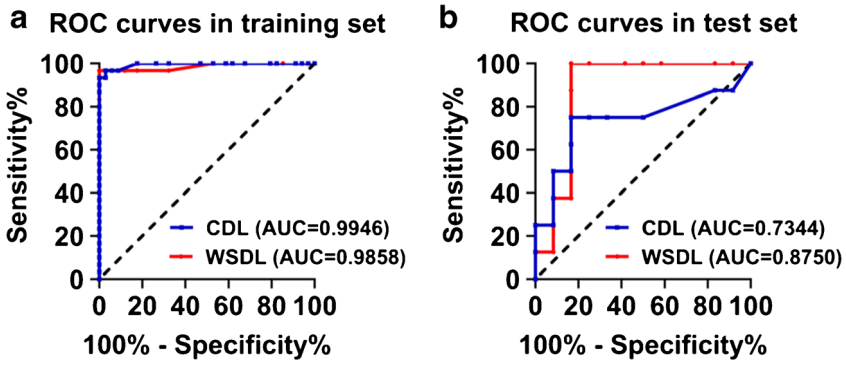


Fig. 4 ROC curves comparing the predictive power of PSI for PFS in the training (A) and test (B) sets. ROC, receiver operator characteristic; AUC, area under the curve; PSI, prediction similarity index; WSDL, weakly supervised deep learning; CDL, conventional deep learning

Table 2 Deep learning feature-based detection efficiency and prognosis prediction

	Training set with WSDL (n=64)	Test set with WSDL (n=20)	Training set with CDL (n=64)	Test set with CDL (n=20)
Sensitivity	86.67%	87.50%	73.33%	62.5%
Specificity	100%	83.33%	100%	83.33%
Accuracy	93.75%	85.00%	87.50%	75.00%
2-year PFS (PSI>1)	34.6%±9.3%	33.3%±15.7%	36.4%±10.3%	28.6%±17.1%
2-year PFS (PSI<1)	92.1%±4.4%	90.9%±8.7%	85.7%±5.4%	84.6%±10.0%
5-year PFS (PSI>1)	3.8%±3.8%	22.2%±13.9%	4.5%±4.4%	28.6%±17.1%
5-year PFS (PSI<1)	92.1%±4.4%	90.9%±8.7%	77.1%±9.5%	74.0%±13.2%

Abbreviations: PFS, progression-free survival; PSI, prediction similarity index; WSDL, weakly supervised deep learning; CDL, conventional deep learning

with different PFS in case of both the WSDL ($P < 0.0001$) and CDL ($P < 0.0001$) methods (Fig. 5A and C). Similarly, in the test set, the WSDL ($P = 0.0017$) and CDL ($P = 0.0177$) methods could distinguish patients with different PFS (Fig. 5B and D).

Predictive value of other clinical and imaging parameters and integrated analysis

Major clinical factors, such as gender, serum lactate dehydrogenase levels, ECOG score, β2-microglobulin levels, and Epstein–Barr virus DNA, were significantly associated with PFS in univariate analysis. Conventional imaging parameters, including PET/CT-based Ann Arbor stage, MTV, and TLG, were also significantly associated with PFS in univariate analysis (refer to Table 3 for more details). Furthermore, we combined PSI with these clinical parameters to analyze the prognosis of ENKTL using the multivariate Cox proportional hazard model. We found that PSI was the only independent significant predictor of PFS. The WSDL method (HR, 15.183; 95% CI, 5.479–42.077; $P = 0.000$) achieved better PFS prognosis than the CDL method (HR, 7.857; 95% CI, 3.276–18.843; $P = 0.000$) after adjustment for various cofactors, as listed above.

Discussion

The prognosis of high-risk ENKTL patients is generally poor [32, 38], and treating such patients is thus challenging. Although new regimes have been proposed, the response remains suboptimal due to strong disease heterogeneity [38]. Prognostic index of natural killer lymphoma (PINK) is a well-established index based on age, serum lactate dehydrogenase level, performance status, and disease stage. The PINK model [39] is based on clinical information; patients with the

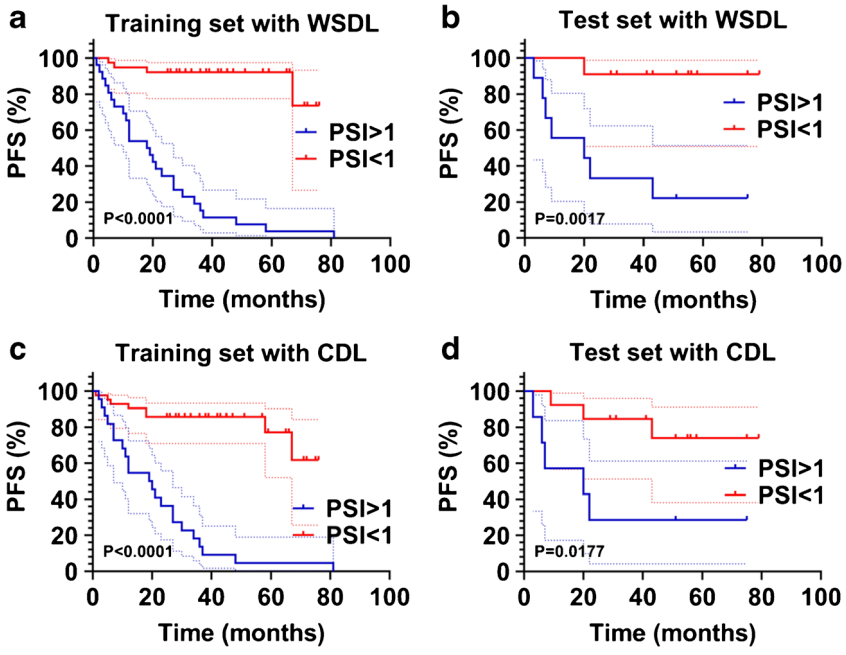


Fig. 5 Kaplan–Meier estimates of PFS in the training (A) and test (B) sets of patients with high and low PSI. PFS, progression-free survival; PSI, prediction similarity index; WSDL, weakly supervised deep learning; CDL, conventional deep learning

Table 3 Univariate analysis involving patients with follow-up data

Characteristics	Training cohort (n=64)		Test cohort (n=20)		Total (n=84)	
	Cutoff value	P	Cutoff value	P	Cutoff value	P
Gender	M/F	0.100	M/F	0.017	M/F	0.010
Age	60	0.184	60	0.041	60	0.742
Serum LDH	169* (0.092)	0.263	223* (0.137)	0.065	231.5 (0.019)	0.000
ECOG score	0/1/2/3/4	0.057	0/1/2/3/4	0.023	0/1/2/3/4	0.005
Ki67	60%* (0.767)	0.548	80%* (0.665)	0.870	70%* (0.970)	0.809
β2-microglobulin	188* (0.076)	0.387	454* (0.248)	0.328	820 (0.040)	0.001
EBV DNA	+/−	0.018	+/−	0.012	+/−	0.001
Ann Arbor stage	I–II/III–IV	0.000	I–II/III–IV	0.000	I–II/III–IV	0.000
B symptoms	+/−	0.300	+/−	0.441	+/−	0.193
PSI with CDL	1	0.000	1	0.018	1	0.000
PSI with WSDL	1	0.000	1	0.002	1	0.000
SUVmax	11.1* (0.382)	0.876	15.05* (0.418)	0.880	12.25* (0.218)	0.871
SUVmean	6.35* (0.453)	0.927	8.6* (0.298)	0.312	6.875* (0.249)	0.677
MTV	18.04 (0.002)	0.000	15.695* (0.165)	0.415	25.325 (0.001)	0.000
TLG	94.738 (0.004)	0.000	124.133* (0.316)	0.415	109.952 (0.006)	0.001

*Median value

Abbreviations: M, male; F, female; +, positive; −, negative; PFS, progression-free survival; PSI, prediction similarity index; WSDL, weakly supervised deep learning; CDL, conventional deep learning; LDH, lactate dehydrogenase; ECOG, Eastern Cooperative Oncology Group; EB virus, Epstein–Barr virus; SUV, standardized uptake value; MTV, metabolic tumor volume; TLG, total lesion glycolysis

same PINK score could even show different prognosis. As a clinical molecular imaging method, ^{18}F-FDG PET/CT shows good potential to help stratify patients and optimize prognosis for the treatment of many types of cancers [9, 40–42]. However, considering the low incidence of ENKTL, the potential of this method for predicting the prognosis of ENKTL remains poorly explored. Conventional ^{18}F-FDG PET/CT-related parameters, such as SUVmax, SUVmean, MTV, and TGL, have been found to show a correlation with survival, but the results have been debatable [30, 31, 36, 43]. These parameters cannot facilitate a comprehensive image-based analysis of tumors and cannot be integrated in hematological guidelines [44] because prospective studies with larger cohort of patients and methodological harmonization are needed [45]. Our univariate analysis indicated that SUVmax and SUVmean were not related to prognosis, while MTV and TGL were related to prognosis. However, multivariate analyses indicated that none of them were associated with prognosis. Considering the rarity of ENKTL, it is difficult to predict its prognosis, particularly in small cohort of patients.

Considering the potential of AI in facilitating data analyses to discover useful information, we aimed to develop and validate AI methods to overcome the restriction of limited data availability and to explore the prognostic value of ^{18}F-FDG PET/CT in ENKTL. We herein proposed an AI model that could utilize incomplete or missing follow-up data to enhance the prediction potential of deep learning methods. This improved prediction power of AI led to the extraction of feature maps from ^{18}F-FDG PET/CT as effective surrogates for prognosis prediction in patients with ENKTL. Furthermore, the method could automatically discover characteristic features in metabolic imaging. Our results confirmed the benefits of AI for comprehensive imaging analyses, wherein the proposed PSI was better than conventional clinical parameters and other PET-related parameters for prognosis prediction.

AI methods tend to be biased toward texture rather than shape, while human cognitive processes function in the opposite manner [46]. Conventional ^{18}F-FDG PET/CT-related parameters, such as Ann Arbor stage, SUVmax, SUVmean, MTV, and TGL, have been already covered within the AI framework, and they reportedly have inferior predictive performance than deep learning methods [47]. The current developments occurring within the field of AI can add value to conventional PET analyses. To avoid redundancy and correlation of tested data and to lower the number of parameters tested in view of the limited size of our cohort, Ann Arbor stage, MTV, and TLG were not included in multivariate analysis, although they were found to be related to prognosis in univariate analysis. For multivariate analysis, clinical prognostic factors and PSI were included. PSI eventually emerged to be the only independent predictor of PFS.

Despite their potential, the application of AI-based methods to clinical trials remains challenging due to limited sample sizes. Deep learning research is particularly difficult

for rare diseases such as ENKTL. Moreover, not all recruited patients can be finally enrolled due to missing or incomplete follow-up. Therefore, we developed a WSDL method in an attempt to solve this problem. During the training of WSDL, implicit labels are generated by exploring similarities among patients, and this diversity can be captured by a deep neural network. Most supervised data augmentation methods have been developed by using unlabeled data for regularization under particular distributional assumptions, such as cluster or smoothness assumption [48]. However, the performance of such a model can be considerably deteriorated if the real data distribution violates the assumed distribution [14]. In this study, the proposed WSDL method with integrated PNU strategy did not make additional assumptions about data distribution; therefore, the performance of prognosis prediction was efficiently and robustly improved. We conducted a pilot study to reutilize the data without follow-up information to boost the prediction accuracy of patient survival; consequently, the advantages of the proposed WSDL method were confirmed in our test set. By employing WSDL, prognoses of patients in the test set could be significantly differentiated, and the results were better than on using CDL. Therefore, the proposed WSDL method may act as a practical tool for developing individualized treatment strategies using clinical trial data.

Tumor heterogeneity in baseline PET/CT images may allow better signature characterization and improve prediction of therapy response and survival in malignant tumors [49, 50]. Ko et al. [49] investigated whether the textural features of pretreatment ^{18}F-FDG PET images could predict the prognosis for ENKTL; they reported that dissimilarity and low-intensity short-zone emphasis were significant predictors of disease progression in patients with ENKTL and were able to improve their prognostic stratification. However, there were only 17 patients in this retrospective study and details pertaining to the regimen were not mentioned. In our study, PSI was validated as a potential index for risk stratification and future management of patients with ENKTL. Compared with texture analyses, the results of deep learning are more difficult to interpret. Deep learning–based radiomics studies [9] evidently draw several image-based texture parameters and the significance of many of them cannot be explained in a clinical perspective; this hinders the application in clinical routine. In addition to the proposed PSI, we also visualized the extracted features as strips of feature maps. Although these maps did not give us an in-depth insight into physiological interpretation, they did give us an additional view of recommendations derived from the black box, and the different activation patterns may facilitate quality control in practice. The feature maps were composed of multiple features, and, therefore, they contained more information than a single scalar value of PSI. An increase in the dimension of the features may improve prediction but may lead to overfitting. On the other hand, a single scalar value is convenient for clinical interpretation. Therefore, it may be practical to consider both PSI values and feature maps to gather better, more robust information.

This study had several limitations. First, although we employed WSDL to enhance data utilization, the sample size was still small, which may reduce the test power and predictive ability of deep learning methods. Similar to other studies based on rare diseases, the difference between overall survival and PFS was not great, and we did not perform overall survival-related survival analysis. We only performed survival analysis based on PFS. Second, tumors were outlined by a specialist in medical radiology and nuclear medicine. As with previous studies, interobserver variations may exist in the manual delineation and may influence the reported results [9]. Nevertheless, deep learning methods can automatically learn features included in the hidden layers of neural networks from imaging data, and they are less sensitive to segmentation variations [51, 52]. Third, study data were collected from a single center, and external validation is thus necessary to validate our findings. Finally, potential patient selection biases may exist because of the retrospective nature of this study.

To summarize, our proposed WDSL method was able to utilize incomplete or missing follow-up data to improve survival prediction. Deep learning involving ^{18}F-FDG PET/CT provides an effective approach for prognosis prediction in patients with ENKTL. The identified feature maps and PSI may potentially assist the stratification of patients in therapy. Future prospective studies with external validation are nevertheless warranted to validate our findings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-021-05232-3.

Funding This work was supported by the National Natural Science Foundation of China (No. 81974276), the 3-year planning of the Shanghai Shen-Kang Promoting Hospital's Clinical Skills and Innovative Ability Project (No. 16CR3110B) and Shanghai Municipal Key Clinical Specialty (No. shslczdzk03403).

Declarations

Ethics approval All procedures involving human participants were performed in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. No experiments involved animals.

Informed consent Informed consent was obtained from all patients included in this study.

Conflict of interest The authors declare no conflict of interest.

References

1. Hatt M, Le Rest CC, Tixier F, Badic B, Schick U, Visvikis D. Radiomics: data are also images. J Nucl Med. 2019;60(Suppl 2): 38S–44S.
2. Visvikis D, Cheze Le Rest C, Jaouen V, Hatt M. Artificial intelligence, machine (deep) learning and radio(geno)mics: definitions and nuclear medicine imaging applications. Eur J Nucl Med Mol Imaging. 2019;46:2630–7.
3. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology. 2020;296(2):E65–71.
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
5. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–31.
6. Blanc-Durand P, Campedel L, Mule S, Jegou S, Luciani A, Pigneur F, et al. Prognostic value of anthropometric measures extracted from whole-body CT using deep learning in patients with non-small-cell lung cancer. Eur Radiol. 2020;30:3528–37.
7. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin Cancer Res. 2019;25:3266–75.
8. Tang Z, Xu Y, Jin L, Aibaidula A, Lu J, Jiao Z, et al. Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients. IEEE Trans Med Imaging. 2020;39: 2100–9.
9. Peng H, Dong D, Fang MJ, Li L, Tang LL, Chen L, et al. Prognostic value of deep learning PET/CT-based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. Clin Cancer Res. 2019;25:4271–9.
10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
11. Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. Classification in BioApps. 2018:323–50.
12. Karimi D, Nir G, Fazli L, Black PC, Goldenberg L, Salcudean SE. Deep learning-based Gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation. IEEE J Biomed Health Inform. 2020;24:1413–26.
13. Leunens G, Verstraete J, Van den Bogaert W, Van Dam J, Dutreix A, van der Schueren E. Human errors in data transfer during the preparation and delivery of radiation treatment affecting the final result: "garbage in, garbage out". Radiother Oncol. 1992;23:217–22.
14. Krijthe JH, Loog M. Robust semi-supervised least squares classification by implicit constraints. Pattern Recogn. 2017;63:115–26.
15. Sakai T, MCdP, Niu G, Sugiyama M. Semi-supervised classification based on classification from positive and unlabeled data. The 34th International Conference on Machine Learning. Sydney, Australia; 2017;2998–3006.
16. Yu-Feng Li Z-HZ. Towards making unlabeled data never hurt. IEEE Trans Pattern Anal Mach Intell. 2015;37:14.
17. Lee J, Suh C, Park YH, Ko YH, Bang SM, Lee JH, et al. Extranodal natural killer T-cell lymphoma, nasal-type: a prognostic model from a retrospective multicenter study. J Clin Oncol. 2006;24:612–8.
18. Au WY, Ma SY, Chim CS, Choy C, Loong F, Lie AK, et al. Clinicopathologic features and treatment outcome of mature T-cell and natural killer-cell lymphomas diagnosed according to the World Health Organization classification scheme: a single center experience of 10 years. Ann Oncol. 2005;16:206–14.
19. Li CC, Tien HF, Tang JL, Yao M, Chen YC, Su IJ, et al. Treatment outcome and pattern of failure in 77 patients with sinonasal natural killer/T-cell or T-cell lymphoma. Cancer. 2004;100:366–75.
20. The world health organization classification of malignant lymphomas in japan: incidence of recently recognized entities. Lymphoma Study Group of Japanese Pathologists. Pathol Int. 2000;50:696–702.
21. Chen CY, Yao M, Tang JL, Tsay W, Wang CC, Chou WC, et al. Chromosomal abnormalities of 200 Chinese patients with non-Hodgkin's lymphoma in Taiwan: with special reference to T-cell lymphoma. Ann Oncol. 2004;15:1091–6.
22. Chan WK, Au WY, Wong CY, Liang R, Leung AY, Kwong YL, et al. Metabolic activity measured by F-18 FDG PET in natural killer-cell lymphoma compared to aggressive B- and T-cell lymphomas. Clin Nucl Med. 2010;35:571–5.
23. Khong PL, Pang CB, Liang R, Kwong YL, Au WY. Fluorine-18 fluorodeoxyglucose positron emission tomography in mature T-cell and natural killer cell malignancies. Ann Hematol. 2008;87:613–21.
24. Moon SH, Cho SK, Kim WS, Kim SJ, Chan Ahn Y, Choe YS, et al. The role of 18F-FDG PET/CT for initial staging of nasal type natural killer/T-cell lymphoma: a comparison with conventional staging methods. J Nucl Med. 2013;54:1039–44.
25. Zhou X, Lu K, Geng L, Li X, Jiang Y, Wang X. Utility of PET/CT in the diagnosis and staging of extranodal natural killer/T-cell lymphoma: a systematic review and meta-analysis. Medicine (Baltimore). 2014;93:e258.
26. Casulo C, Schoder H, Feeney J, Lim R, Maragulia J, Zelenetz AD, et al. 18F-fluorodeoxyglucose positron emission tomography in the staging and prognosis of T cell lymphoma. Leuk Lymphoma. 2013;54:2163–7.
27. Fujiwara H, Maeda Y, Nawa Y, Yamakura M, Ennishi D, Miyazaki Y, et al. The utility of positron emission tomography/computed tomography in the staging of extranodal natural killer/T-cell lymphoma. Eur J Haematol. 2011;87:123–9.
28. Wu HB, Wang QS, Wang MF, Li HS, Zhou WL, Ye XH, et al. Utility of 18F-FDG PET/CT for staging NK/T-cell lymphomas. Nucl Med Commun. 2010;31:195–200.
29. Karantanis D, Subramaniam RM, Peller PJ, Lowe VJ, Durski JM, Collins DA, et al. The value of [(18)F]fluorodeoxyglucose positron emission tomography/computed tomography in extranodal natural killer/T-cell lymphoma. Clin Lymphoma Myeloma. 2008;8:94–9.
30. Suh C, Kang YK, Roh JL, Kim MR, Kim JS, Huh J, et al. Prognostic value of tumor 18F-FDG uptake in patients with untreated extranodal natural killer/T-cell lymphomas of the head and neck. J Nucl Med. 2008;49:1783–9.
31. Khong PL, Huang B, Lee EY, Chan WK, Kwong YL. Midtreatment 18F-FDG PET/CT scan for early response assessment of SMILE therapy in natural killer/T-cell lymphoma: a prospective study from a single center. J Nucl Med. 2014;55:911–6.

32. Guo R, Xu P, Xu H, Miao Y, Li B. The predictive value of pretreatment 18F-FDG PET/CT on treatment outcome in early-stage extranodal natural killer/T-cell lymphoma. Leuk Lymphoma. 2020;61(11):2659–64.
33. Bai B, Huang HQ, Cai QC, Fan W, Wang XX, Zhang X, et al. Predictive value of pretreatment positron emission tomography/computed tomography in patients with newly diagnosed extranodal natural killer/T-cell lymphoma. Med Oncol. 2013;30:339.
34. Chang Y, Fu X, Sun Z, Xie X, Wang R, Li Z, et al. Utility of baseline, interim and end-of-treatment (18)F-FDG PET/CT in extranodal natural killer/T-cell lymphoma patients treated with L-asparaginase/pegaspargase. Sci Rep. 2017;7:41057.
35. Jiang C, Zhang X, Jiang M, Zou L, Su M, Kosik RO, et al. Assessment of the prognostic capacity of pretreatment, interim, and post-therapy (18)F-FDG PET/CT in extranodal natural killer/T-cell lymphoma, nasal type. Ann Nucl Med. 2015;29:442–51.
36. Jiang C, Su M, Kosik RO, Zou L, Jiang M, Tian R. The Deauville 5-point scale improves the prognostic value of interim FDG PET/CT in extranodal natural killer/T-cell lymphoma. Clin Nucl Med. 2015;40:767–73.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016;770–8.
38. Tse E, Kwong YL. The diagnosis and management of NK/T-cell lymphomas. J Hematol Oncol. 2017;10:85.
39. Kim SJ, Yoon DH, Jaccard A, Chng WJ, Lim ST, Hong H, et al. A prognostic index for natural killer cell lymphoma after non-anthracycline-based treatment: a multicentre, retrospective analysis. Lancet Oncol. 2016;17:389–400.
40. Cheng NM, Hsieh CE, Fang YD, Liao CT, Ng SH, Wang HM, et al. Development and validation of a prognostic model incorporating [(18)F]FDG PET/CT radiomics for patients with minor salivary gland carcinoma. EJNMMI Res. 2020;10:74.
41. Senjo H, Hirata K, Izumiyama K, Minauchi K, Tsukamoto E, Itoh K, et al. High metabolic heterogeneity on baseline 18FDG-PET/CT scan as a poor prognostic factor for newly diagnosed diffuse large B-cell lymphoma. Blood Adv. 2020;4:2286–96.
42. Pinho DF, King B, Xi Y, Albuquerque K, Lea J, Subramaniam RM. Value of Intratumoral metabolic heterogeneity and quantitative (18)F-FDG PET/CT parameters in predicting prognosis for patients with cervical cancer. AJR Am J Roentgenol. 2020;214:908–16.
43. Kim CY, Hong CM, Kim DH, Son SH, Jeong SY, Lee SW, et al. Prognostic value of whole-body metabolic tumour volume and total lesion glycolysis measured on (18)F-FDG PET/CT in patients with extranodal NK/T-cell lymphoma. Eur J Nucl Med Mol Imaging. 2013;40:1321–9.
44. Barrington SF, Mikhaeel NG, Kostakoglu L, Meignan M, Hutchings M, Mueller SP, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. J Clin Oncol. 2014;32:3048–58.
45. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. Eur J Nucl Med Mol Imaging. 2017;44:17–31.
46. Kubilius J, Bracci S, Op de Beeck HP. Deep neural networks as a computational model for human shape sensitivity. PLoS Comput Biol. 2016;12:e1004896.
47. Baek S, He Y, Allen BG, Buatti JM, Smith BJ, Tong L, et al. Deep segmentation networks predict survival of non-small cell lung cancer. Sci Rep. 2019;9:17286.
48. Chapelle O, Zien A. Semi-supervised classification by low density separation. AISTATS. 2005:57–64.
49. Ko KY, Liu CJ, Ko CL, Yen RF. Intratumoral heterogeneity of pretreatment 18F-FDG PET images predict disease progression in patients with nasal type extranodal natural killer/T-cell lymphoma. Clin Nucl Med. 2016;41:922–6.
50. Gao J, Huang X, Meng H, Zhang M, Zhang X, Lin X, et al. Performance of multiparametric functional imaging and texture analysis in predicting synchronous metastatic disease in pancreatic ductal adenocarcinoma patients by hybrid PET/MR: initial experience. Front Oncol. 2020;10:198.
51. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
52. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging. 2016;35:1299–312.

6

Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features

This chapter has been published as **peer-reviewed conference paper:**.

X. Hu, Y. Yan, W. Ren, H. Li, A. Bayat, Y. Zhao, and B. Menze. "Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features." In: *Medical Imaging with Deep Learning*. PMLR. 2021

Synopsis: This work proposes a Feedback Graph Attention Convolutional Network (FB-GACN) for MR image enhancement via a self-similarity learning strategy to update the features of each node in a graph. Learning the symmetry and similarity relationship of each pair, the content with the same texture (e.g., edges, corners, and lesions) gets sharper and can be used to remove some artifacts. It recovers more texture details by employing the feedback mechanism (consecutive iterations) to facilitate low-resolution (LR) images to reconstruct super-resolution (SR) images. The proposed network achieves better high-resolution criteria and superior visual quality compared to state-of-the-art methods in two crucial tasks: i) cross-protocol super resolution of diffusion MRI and ii) MRI artifacts removal.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features

Xiaobin Hu[*1] XIAOBIN.HU@TUM.DE
Yanyang Yan[*2] YANYANYANG@IIE.AC.CN
Wenqi Ren[2] RWQ.RENWENQI@GMAIL.COM
Hongwei Li[1] HONGWEI.LI@TUM.DE
Amirhossein Bayat[1] AMIR.BAYAT@TUM.DE
Yu Zhao[1] YUZHAO90@OUTLOOK.COM
Bjoern Menze[1] BJOERN.MENZE@TUM.DE
[1] *Department of Computer Science, Technische Universität München, Munich*
[2] *Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China*

Abstract

Artifacts, blur, and noise are the common distortions degrading MRI images during the acquisition process, and deep neural networks have been demonstrated to help in improving image quality. To well exploit global structural information and self-similarity details, we propose a novel MR image enhancement network, named Feedback Graph Attention Convolutional Network (FB-GACN). As a key innovation, we consider the global structure of an image by building a graph network from image sub-regions that we consider to be node features, linking them non-locally according to their similarity. The proposed model consists of three main parts: 1) The parallel graph similarity branch and content branch, where the graph similarity branch aims at exploiting the similarity and symmetry across different image sub-regions in low-resolution feature space and provides additional priors for the content branch to enhance texture details. 2) A feedback mechanism with a recurrent structure to refine low-level representations with high-level information and generate powerful high-level texture details by handling the feedback connections. 3) A reconstruction to remove the artifacts and recover super-resolution images by using the estimated sub-region self-similarity priors obtained from the graph similarity branch. We evaluate our method on two image enhancement tasks: i) cross-protocol super resolution of diffusion MRI; ii) artifact removal of FLAIR MR images. Experimental results demonstrate that the proposed algorithm outperforms the state-of-the-art methods.

Keywords: Magnetic resonance imaging, image enhancement, self-similarity, graph similarity branch, feedback mechanism.

1. Introduction

For Magnetic Resonance Imaging (MRI) sequences, it is an inevitable dilemma to achieve a balance between image resolution, signal-to-noise ratio, and acquisition time (Brown et al., 2014). Higher resolution imaging grasps more structural details and provides more diagnostic information, but requires longer acquisition time (Sui et al., 2019). Since the signal-to-noise ratio is proportional to the slice thickness and the square root of scanning time, the longer acquisition time leads to the performance drop of the signal-to-noise ratio and tends

[*] Contributed equally

to generate artifacts caused by physiologic motion such as respiratory motion and physical movement of subjects. Considering the limited and costly MRI resource, some thick slices and low scan time MRI images are usually utilized to get a desired signal-to-noise ratio (Lee et al., 2020; Wu et al., 2019; Meurée et al., 2019). Consequently, the use of image enhancement techniques is an established field of research in medical image computing and imaging physics (Shi et al., 2015), for example, to prevent blurring and information loss when co-aligning different image volumes in a multi-parametric sequence.

Recently, Convolutional Neural Network (CNN) based approaches have shown dramatic improvements over traditional super-resolution (SR) methods and exhibited state-of-the-art performance in natural and medical images. A super-resolution convolutional neural network (SRCNN) (Dong et al., 2014) was proposed to learn a nonlinear mapping between the low-resolution (LR) and high-resolution (HR) images. Wide residual networks with fixed skip connections (Shi et al., 2018) was presented for MR images super-resolution. A new CNN-based model (Tanno et al., 2017) was proposed for a diffusion tensor imaging SR task. Besides, Graph Neural Networks (GNN) have also shown their powerful ability to exploit structural information dealing with data of graph structure. The notation of GNN was firstly introduced (Gori et al., 2005), and then further elaborated as a generalization of recursive neural networks, which is widely used to explore the structural characters in various applications including chemistry, recommender systems, and social network study to deal with challenge tasks, e.g., finding the chemical compounds that are most similar to a query compound, tackling the graph similarity computation for query systems (Bai et al., 2019). Nowadays, it is an interesting trend to combine GNN and CNN to develop their corresponding advantages (Veličković et al., 2018). GNNs help with reducing the data dimensionality from image features extracted by CNN to high-level and compact features in graph nodes. FCNs are limited in the receptive field. Adding a GNNs could increase the receptive field of networks when dealing with large images. The combination of CCN and GNN is a convolutional graph neural network that generalizes the operation of convolution from grid data to graph data. It plays a central role in building up many complex GNN models (Wu et al., 2020).

To avoid generating inconsistent HR results after replacing the LR patches, in our method, the similar patch pairs are matched in feature space and the graph attention mechanism is used to update features representation of each patch (node) with the adaptive weight combination of those similar patches' features. As far as we know, it is the first work to explore the self-similarity and continuous relationship of MRI and fully exploit the feedback mechanism to increase the reconstruction accuracy for MR images. More specifically, in this paper, we propose a novel biomedical image enhancement network based on the feedback mechanism and graph attention convolutional network, where graph networks are employed as a self-similarity strategy which assigns larger weights to the more important and similar nodes or features.

The main contributions of this paper are:

1) We propose a Feedback Graph Attention Convolutional Network (FB-GACN) for MR image enhancement. To the best of our knowledge, it is the first work to construct a graph-based network into the image enhancement by exploring globally structural similarity among similar paired sub-regions.

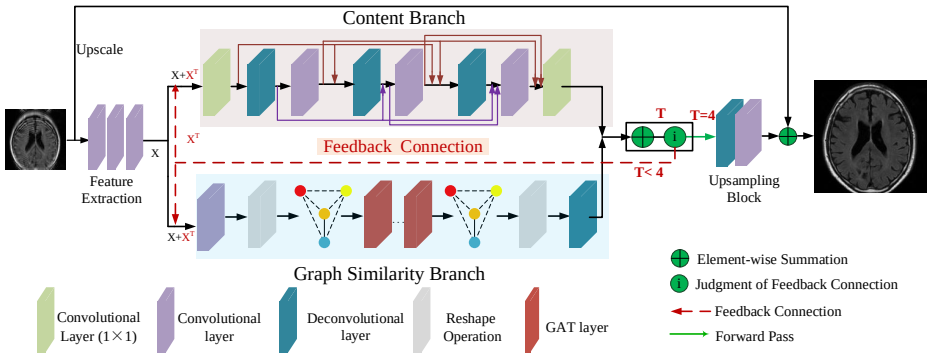


Figure 1: Architecture of the proposed FB-GACN model. Our FB-GACN contains three parts: 1) The content block to generate the high-level texture details. 2) The graph attention branch to exploit the similarity and symmetric knowledge across MRI patches. 3) A reconstruction to remove the artifact and reconstruct super-resolution MRI by using the estimated patch correlation priors. The feedback mechanism is the recurrent structure to refine x features with high-level x^T by the feedback connections.

2) We propose a self-similarity learning strategy to update the features of each node in a graph. Learning the symmetry and similarity relationship of each pair, the content with same texture (e.g., edges, corners, and lesions) gets sharper and can be used to remove some artifacts. It recovers more texture details by employing the feedback mechanism (consecutive iterations) to facilitate LR images to reconstruct SR images.

3) We demonstrate the performance in two crucial tasks: i) cross-protocol super resolution of diffusion MRI and ii) MRI artifacts removal. The proposed network achieves better high-resolution criteria and superior visual quality compared to state-of-the-art methods.

2. Method

The whole pipeline consists of following three steps. Firstly, a stack of convolution layers extracts the low-resolution features of input distortion images. Afterward, the content branch and graph similarity branch work parallel to exploit the texture and self-similarity information. Finally, the upsampling block reconstructs final super-resolution results using the estimated patch correlation and texture priors.

Specialized design for MR images: Our method aims to learn the symmetry and self-similarity relationship of patch-based features in multi-modal brain MR images where the structure of the brain is normally symmetry, shown in Fig. 2 (a). To meet this requirement, we designed a specialized Graph-based structure to merge the high-similarity information of sub-regions by updating larger weights to the more important and similar nodes or features in a graph attention fashion.

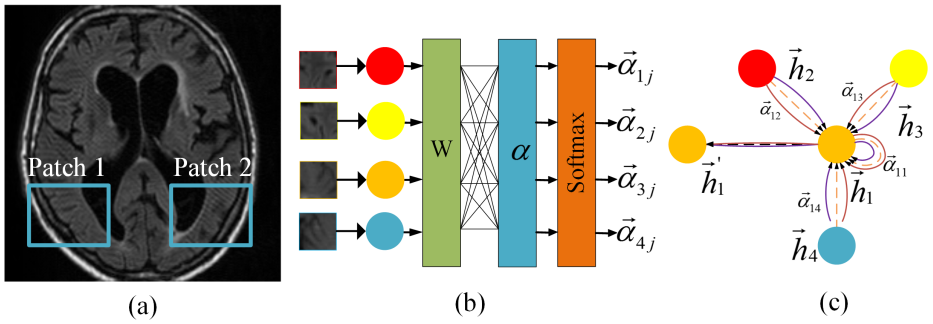


Figure 2: **(a)** Exploring the self-similarity features to remove artifacts: Swapping the artifacts features in Patch 2 with clear features of Patch 1. **(b)** The employed attention mechanism. A shared linear transformation W is applied to every node. Afterwards, a self-attention mechanism a is calculated on features to learn the correlation among nodes. **(c)** An illustration of multi-head attention mechanism by node 1 on its neighbors.

2.1. Architecture of FB-GACN

The structure of the proposed FB-GACN is illustrated in Fig. 1. A long skip connection is added to pass the upsampled LR image to the output result as we only want to learn the residual modifications. After feature extraction, the output are low-resolution features with the dimension of $h \times w \times d$, where h and w denote the spatial dimension of the LR input and d is the number of feature channels. Then the LR features are imported into the content branch and graph similarity branch, respectively. The upsampling block U is made up of deconvolution layers to upscale the HR features, and convolutional layers to recover a residual image. The final reconstruction SR images are the pixel-wise sum of the upsampled LR input and the residual image. The mathematical formulation is elaborated as:

$$I^{SR} = f_U\left[f_G\left(f_E\left(I^{LR}\right)\right) + f_F\left(f_E\left(I^{LR}\right)\right)\right] + I^{LR}_{up}, \tag{1}$$

where $f_E(\cdot)$, $f_G(\cdot)$, $f_F(\cdot)$, and $f_U(\cdot)$ represent the operations of the feature extraction E, graph similarity branch G, content branch F and upsamling U blocks, respectively. The objective function is L_1 norm-based loss function. The network is trained by minimizing the objective function as following:

$$\ell_{(\theta)} = \frac{1}{n}\sum_{i=1}^{n}\left\|I_i^{SR} - I_i^{HR}\right\|_1, \tag{2}$$

whre θ and n are the parameters of the network and the number of images pairs, respectively. I_i^{SR} is the reconstruction of super-resolution MRI, and I_i^{HR} is the corresponding ground truth.

2.2. Graph Similarity Branch

Graph similarity branch employs graph attention network layers (GAT) (Veličković et al., 2018) to make use of the contextual information among image patches to help recover structure and remove artifacts. After feeding the extracted LR feature maps to a convolutional layer with stride of s and kernel size of p, we form a graph using the $n \times d$ matrix where

we assume there exist n nodes with d-th dimensional features. Each node is connected with five neighboring nodes and the attention coefficient of each node is updated. The single graph attention layer is shown in Fig. 2. The input of the single attention layer is a set of node features, $\boldsymbol{h} = \{\overrightarrow{h}_1, \overrightarrow{h}_2, ..., \overrightarrow{h}_N\}$, $h_i \in \mathbb{R}^F$, where N is the number of nodes, and F is the number of features in each node. The GAT layer updates a new set of node features, $\boldsymbol{h}' = \{\overrightarrow{h}'_1, \overrightarrow{h}'_2, ..., \overrightarrow{h}'_N\}$, $h'_i \in \mathbb{R}^{F'}$. Then a learnable linear transformation and self-attention is performed on the nodes (a shared attention mechanism $a : R^{F'} \times R^{F'} \to R^F$ computes attention coefficients):

$$e_{ij} = a(\boldsymbol{W}\overrightarrow{h}_i, \boldsymbol{W}\overrightarrow{h}_j), \tag{3}$$

which represents the importance of node j to node i. Afterwards, the attention coefficients are normalized by the softmax function:

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N}} \exp(e_{ij}), \tag{4}$$

Following (Veličković et al., 2018), the attention mechanism a is a single-layer feedforward neural network, parametrized by weight matrix $\overrightarrow{\boldsymbol{a}} \in \mathbb{R}^{2F'}$. After applying the LeakyReLU nonlinearity, the coefficients are also expressed as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\overrightarrow{a}^T[\boldsymbol{W}\overrightarrow{h}_i \| \boldsymbol{W}\overrightarrow{h}_j]))}{\sum_{k \in N_i} \exp(\text{LeakyReLU}(\overrightarrow{a}^T[\boldsymbol{W}\overrightarrow{h}_i \| \boldsymbol{W}\overrightarrow{h}_k]))}, \tag{5}$$

where $(\cdot)^T$ represents the transposition operations and $\|$ means the concatenation. Then the final output of each node is updated on the strength of the similar neighborhood LR feature nodes $\overrightarrow{h}_j$:

$$\overrightarrow{h}'_i = \sigma\left(\sum_{j \in N} \alpha_{ij} \boldsymbol{W}\overrightarrow{h}_j\right), \tag{6}$$

We also employ the content branch to recover texture details shown in Fig. 1, which is a stack of 3 deconvlutional and 3 convolutional layers.

2.3. Feedback Mechanism

The feedback mechanism is a loop iteration to allow the network to correct previous states and regenerate high-level representations. Such iterative cause-and-effect process helps to achieve the principle of the feedback scheme for image SR: high-level information can guide an LR image to recover a better SR image (Li et al., 2019). In our network, we utilize the feedback mechanism to transfer the feature summation with high-level information got from two branches to the low-level information of an input x. The judgment of the feedback connection controller (shown in Fig. 1) determines the time (T) of the feedback iteration, also named the feedback connection. The feedback mechanism is the recurrent CNN structure to refine x features with high-level x^T by the feedback connections ($T - th$ iteration). It can be unfolded to T iteration, in which each iteration t is temporally ordered from 1 to T. The hidden state of each iteration is tied with the loss function and the weight parameters of each iteration are shared. The input of t-th iteration receives the feedback information t-1 iteration to correct original low-level inputs.

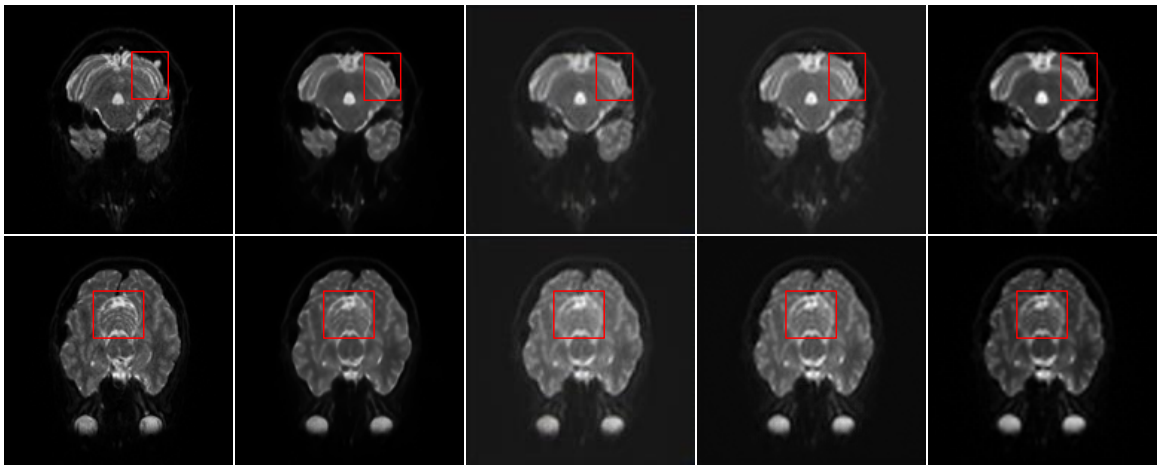
Table 1: Quantitative results of cross-protocol super-resolution and artifacts removal tasks. The best results are highlighted in bold.

Methods	Super-Resolution		Artifacts Removal	
	PSNR	SSIM	PSNR	SSIM
Bicubic	27.34±1.32	0.8882±0.0232	22.58±3.59	0.6855±0.1345
SRCNN (Dong et al., 2014)	29.46±1.68	0.9042±0.0796	24.68±3.38	0.7294±0.1216
VDSR (Kim et al., 2016)	29.66±1.18	0.9026±0.0731	25.39±2.72	0.7588±0.0921
EDSR (Lim et al., 2017)	30.23±1.56	0.9145±0.0229	25.68±3.61	0.7824±0.0952
DDBPN (Haris et al., 2018)	30.34±1.56	0.9171±0.0208	25.58±3.56	0.7821±0.0952
FB-GACN (Ours)	**30.48**±1.63	**0.9185**±0.0194	**25.78**±3.71	**0.7839**±0.1003

3. Experimental Results

3.1. Datasets

Two experiments were conducted to evaluate the performance of the feedback graph attention convolutional network. The first experiment is solving a cross-protocol super-resolution problem on diffusion MRI data (MUSHAC) (Tax et al., 2019). The HR images were obtained by state-of-the-art diffusion MRI acquisition by Prisma scanner with voxel size (1.5 $\times$ 1.5 $\times$ 1.5 mm^3), and the corresponding LR images were scanned by the standard acquisition of Prisma with a larger voxel size (2.4 $\times$ 2.4 $\times$ 2.4 mm^3). Nine subjects are used as training set and one subject for testing. For the second experiment, we utilize the proposed network to remove the MRI artifacts and regenerate HR images by the scale $\times$2. We randomly divided the public WMH dataset (Kuijf et al., 2019) into training (2225 images from 48 patients), validation (278 images from 6 patients) and test parts (278 images from 6 patients). Afterward, the simulated artifacts of FLAIR modality (Kuijf et al., 2019) were generated by the physical model of MRI motion artifacts.



Ground truth Proposed VDSR SRCNN Bicubic

Figure 3: Comparison with state-of-the-art methods of cross-protocol super-resolution on the diffusion MRI data (MUSHAC). Best viewed by zooming in on the screen.

3.2. Implementation Details

In each training batch, nine LR patches are randomly extracted as inputs. We train our model 300 epochs with ADAM optimize and learning rate is set as 10^{-4} initially and is divided by 2 every 80 epochs. We implement experiments with PyTorch using a NVIDIA TITAN X GPU.

3.3. Comparisons with State-of-the-Art Methods

In order to evaluate the performances of our algorithms, we compare them with the start-of-the-art methods qualitatively and quantitatively. The four most recent state-of-the-art super-resolution methods are listed as follows: the Very Deep Super Resolution Network (VDSR) from (Kim et al., 2016), the Super-Resolution Convolutional Neural Network (SRCNN) from (Dong et al., 2014),the Enhanced Deep Residual Networks (EDSR) from (Lim et al., 2017), and the Deep Back-Projection Networks For Super-Resolution (DBPN) from (Haris et al., 2018). We use open-resource implementations from the authors and train all the networks on the same dataset for a fair comparison.

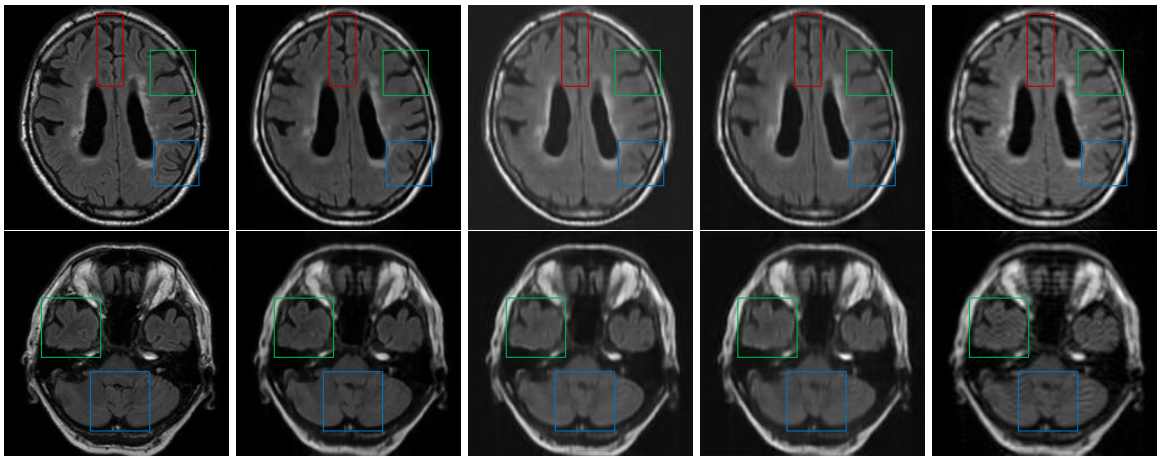


Figure 4: Comparison with state-of-the-art methods of artifacts removal with magnification factors $\times 2$ and the input size 100×100. Best viewed by zooming in on the screen.

3.4. Quantitative Results

The quantitative evaluation of the network using the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) scores are listed in Table 1.

Cross-Protocol Super-Resolution: This task is to evaluate the the performance of our method on the cross-protocol diffusion MRI quality enhancement. Our method achieves better results in comparison with other state-of-the-art methods, especially 3.46 dB higher than the traditional bicubic interpolation method.

Artifacts Removal: To verify the effectiveness of our proposed network towards removing MRI artifacts and super-resolution scale $\times 2$, the PSNR and SSIM results of MRI artifacts are listed in Table 1. Our method outperforms all the state-of-the-art algorithms with the best PSNR 25.78 dB and SSIM 0.7839.

3.5. Qualitative Evaluation

Cross-Protocol Super-Resolution: The qualitative results of our methods on the diffusion MRI data (MUSHAC) by the standard and the start-of-the-art acquisition of Prisma are shown in Figure 3. It can be observed that our proposed method obtains higher visual quality and recovers clearer structures with finer contrast.

Artifacts Removal: The qualitative results of our methods at magnifications $\times 2$ with artifacts are shown in Figure 4. It can be observed that our proposed method can remove artifacts and obtain the super-resolution results from the LR images. It recovers clearer structures with finer contrast, edges and lesion information.

3.6. Ablation study

Table 2: Ablation study results (PSNR/SSIM): Comparisons our proposed model with the configuration without (w/o) the graph similarity knowledge.

Ablation configuration	Super-Resolution	Artifacts Removal
w/o graph similarity	30.35/0.9177	25.65/0.7735
ours	30.48/0.9185	25.77/0.7835

Graph similarity knowledge: We conduct an ablation study to demonstrate the effectiveness of the graph similarity branch. We compare the proposed network with and without patch-based similarity knowledge in terms of PSNR and SSIM on the test data, shown in Table 2. The graph similarity branch boosts the performance both in the super-resolution and artifacts removal tasks.

Feedback Mechanism: We explore the effect of the iterative number of feedback connections. It can be observed from Table 3 that the reconstruction performance is improved when the iterative number increases from $T = 1$ to $T = 4$. Considering the balance between the computational time and the performance, $T = 4$ is chosen as the iterative number in our paper.

Table 3: The impact of the iterative number T of feedback connection.

Feedback Connection	T=1	T=2	T=3	T=4
Super-Resolution	30.22/0.9172	30.28/0.9173	30.34/0.9177	30.48/0.9185
Artifacts Removal	25.26/0.7632	25.41/0.7647	25.49/0.7682	25.77/0.7835

4. Conclusion

In this paper, we proposed a novel feedback graph attention convolutional network to enhance the visual quality and remove the common distortions (e.g., artifacts) of MR images, considering the self-similarity and correlations across MRI sub-regions. We regard each sub-region as a node and construct a graph to capture the global structure. We employ the feedback mechanism to recover texture details by refining low-level representations with high-level information in a time-series way. Comprehensive qualitative and quantitative experiments show that our algorithm can remove artifacts and further generate high-resolution MRI with finer structure, contrast and lesion information.

References

Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, and Wei Wang. Simgnn: A neural network approach to fast graph similarity computation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 384–392, 2019.

Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

Joonhyung Lee, Hyunjong Kim, HyungJin Chung, and Jong Chul Ye. Deep learning fast mri using channel attention in magnitude domain. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 917–920. IEEE, 2020.

Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

Cédric Meurée, Pierre Maurel, Jean-Christophe Ferré, and Christian Barillot. Patch-based super-resolution of arterial spin labeling magnetic resonance images. *Neuroimage*, 189: 85–94, 2019.

Feng Shi, Jian Cheng, Li Wang, Pew-Thian Yap, and Dinggang Shen. Lrtv: Mr image super-resolution with low-rank and total variation regularizations. *IEEE transactions on medical imaging*, 34(12):2459–2466, 2015.

Jun Shi, Zheng Li, Shihui Ying, Chaofeng Wang, Qingping Liu, Qi Zhang, and Pingkun Yan. Mr image super-resolution via wide residual networks with fixed skip connection. *IEEE journal of biomedical and health informatics*, 23(3):1129–1140, 2018.

Yao Sui, Onur Afacan, Ali Gholipour, and Simon K Warfield. Isotropic mri super-resolution reconstruction with multi-scale gradient field prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer, 2019.

Ryutaro Tanno, Daniel E Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.

Chantal MW Tax, Francesco Grussu, Enrico Kaden, Lipeng Ning, Umesh Rudrapatna, C John Evans, Samuel St-Jean, Alexander Leemans, Simon Koppers, Dorit Merhof, et al. Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms. *NeuroImage*, 195:285–299, 2019.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *In: Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

Yan Wu, Yajun Ma, Jing Liu, Jiang Du, and Lei Xing. Self-attention convolutional neural network for improved mr image reconstruction. *Information sciences*, 490:317–328, 2019.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

7

Concluding Remarks

Nowadays, radiologists undertake an increasing number of clinical decisions based on complex readings including medical images and tabular data. This complicated process is a barrier that preventing the physicians from giving an appropriate report in time. Nevertheless, the emergency of deep learning has boosted the development of medical computer-aided technique and have assisted physicians in providing a more exact diagnosis in a shorter time in the clinical workflow by exhibiting a quantitative analysis of suspicious lesions. Furthermore, as a combination of diagnosis, treatment, and prognosis considering individual variability, the success of precision medicine mainly attribute to robust quantitative imaging biomarkers, which can be implemented by deep learning algorithms. In this thesis, we aim to tackle some challenging problems on medical image analysis(i.e., segmentation, prognostic analysis, medical image synthesis) with state-of-the-art deep learning algorithms. Since this thesis is a publication-based paper and each chapter 3 to 6 is independent, this final conclusion provides a summary and a general discussion of medical image analysis as well as outlook in the computer-aided medical system.

7.1 Conclusion

In Chapter 3 we applied the technique of multi-level activation to the nested classes segmentation of glioma. The results of our experiments indicate that the multi-level activation function and its corresponding loss function are efficient compared to Softmax output layer based on the same network framework. Using the MCE loss function and a reweighting scheme with power-law = 0.4, we obtain Dice scores 86% for complete tumor, 77% for tumor core and 72% for enhancing core on the validation leaderboard of the 2018 BRATS challenge, proving the applicability of the multi-level activation scheme. Finally, this activation could be combined with other network architectures. Using it with the best performing architecture of the BRATS challenge

could even lead to further improved results.

In Chapter 4, we presented a coarse-to-fine adversarial network architecture that introduces a multi-scale feature loss to stabilize the adversarial network for biomedical image segmentation tasks. The proposed architecture saves a significant amount of computational memory and time by extracting the lesion bounding box in the coarse stage. Moreover, the results of the four metrics in the experiments (Dice Similarity Coefficient, Sensitivity, Hausdorff Distance and Average Volume Difference) on the ENKL dataset demonstrate that the proposed method outperforms the state-of-the-art U-net segmentation method. The presented method is a general framework and has the potential to be used in general semantic segmentation tasks. We proposed the zone-based uncertainty criteria (lesion-based and background-based criteria) based on Monte Carlo dropout method. The uncertainty of our deep learning model is quantitatively analyzed, and makes it possible to clearly understand the main uncertainty source. Moreover, the quantitative uncertainty analysis provides clinicians with information permitting them to quickly assess, whether they should accept or reject lesions of high uncertainty.

Additionally, in Chapter 5, our proposed WDSL method was able to utilize incomplete or missing follow-up data to improve survival prediction. Deep learning involving 18F-FDG PET/CT provides an effective approach for prognosis prediction in patients with ENKTL. The identified feature maps and PSI may potentially assist the stratification of patients in therapy. Future prospective studies with external validation are nevertheless warranted to validate our findings.

Finally, in Chapter 6, we proposed a novel feedback graph attention convolutional network to enhance the visual quality and remove the common distortions (e.g., artifacts) of MR images, considering the self-similarity and correlations across MRI sub-regions. We regard each sub-region as a node and construct a graph to capture the global structure. We employ the feedback mechanism to recover texture details by refining low-level representations with high-level information in a time-series way. Comprehensive qualitative and quantitative experiments show that our algorithm can remove artifacts and further generate high-resolution MRI with finer structure, contrast and lesion information

7.2 Outlook

7.2.1 Interpretability of Deep Learning

Although there are plenty of successful applications of deep learning models on medical image analysis, the lack of interpretability is a key factor that preventing deep

learning models from widely being adopted in healthcare [6]. This is mainly because physicians find the deep learning models are complicated with numerous unexplained hyper-parameters that are often designed on specific diseases. Considering that the higher interpretability of models can give easier comprehensions and explanations of healthcare decisions for end-users, there is a great desire to make deep learning models more transparent and interpretable. Further, these interpretable deep learning models allow radiologists to design reasonable solutions to make personalized decisions or permit them to quickly assess whether they should accept or reject solutions. Several strategies have been proposed to evaluate and visualize the saliency features of intermediate layers in convolutional networks, such as deconvolution networks [132], guided back-propagation [133] or deep Taylor composition [134], which aim to understand what a network perceives. Another way to improve model transparency is to provide uncertainty or probability analysis for outputs. As a powerful tool for uncertainty, Bayesian deep learning [135] combined Bayesian statistics with deep learning to approximate network uncertainty. Overall, the current interpretation methods have their limitations [136, 137, 138, 139], which fail to give an accurate and comprehensive explanation, and are still an interesting area to be explored.

7.2.2 Neural Architecture Search

For existing deep learning architectures of medical image analysis, the model design heavily relies on the experience of artificial intelligence (AI) researchers and needs higher requirements for radiologists who only have medical background and lack of computer-aided system knowledge. Hence, in order to further decrease the human intervention on model design, neural architecture search is proposed to achieve an end-to-end automatic architecture design. Recently, neural architecture search (NAS) attracts massive interests in developing algorithms to automatically search desirable neural architectures [140, 141, 142, 143, 144] by using search strategies (e.g., evolutionary algorthim [143], reinforcement learning [141] or gradient-based differentiable methods [142]). Architectures obtained by the NAS have achieved highly competitive performance especially in high-level medical image analysis tasks such as image classification [145], localization [146], radiomics [147] and semantic segmentation [148, 149]. But studies of neural architecture search on low-level tasks (e.g., MRI cross-modality synthesis, CT/PET super-resolution) are still limited and can be performed to meet the demand of purely automatic building of neural architecture without manual adjustment.

7.2.3 Federated Learning

Medical data collection of cross-institution and cross-centers is a good approach to build a satisfactory data-driven deep learning architecture within the constraints of data protection and privacy. Since data privacy and protection are crucially important for medical data analysis [150], new techniques also named federated learning [151] are proposed for training models without exposing the underlying training data to the model users. Specifically, each local client learns the local model from its own data and is blind to other client data. The federated learning aggregates the model parameters of other clients at the central server to build up a global model. Although some pilot progress has been achieved on segmentation tasks [152, 153], there are a lot of issues that should be further addressed (e.g., improve model generalizability onto unseen domains).

Appendices

A

List of Publications

The following publications were written *during this thesis*.

Peer-reviewed Journal Articles

- R. Guo*, **X, Hu***, H. Song*, P. Xu, H. Xu, A. Rominger, X. Lin, B. Menze, B. Li, and K. Shi. "Weakly supervised deep learning for determining the prognostic value of 18 F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type." In: *European journal of nuclear medicine and molecular imaging* (2021), pp. 1–11.
- **X. Hu**, R. Guo, J. Chen, H. Li, D. Waldmannstetter, Y. Zhao, B. Li, K. Shi, and B. Menze. "Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images." In: *IEEE journal of biomedical and health informatics* 24.9 (2020), pp. 2599–2608.
- Y. Zhao, H. Li, S. Wan, A. Sekuboyina, **X. Hu**, G. Tetteh, M. Piraud, and B. Menze. "Knowledge-aided convolutional neural network for small organ segmentation." In: *IEEE journal of biomedical and health informatics* 23.4 (2019), pp. 1363–1373. DOI: 10.1109/JBHI.2019.2891526.
- J. Chen, J. Chen, R. Zhang, and **X, Hu**†. "Toward a Brain-Inspired System: Deep Recurrent Reinforcement Learning for a Simulated Self-Driving Agent." In: *Frontiers in neurorobotics* 13 (2019), p. 40.
- C. Ding, **X. Hu**, X. Cui, G. Li, Y. Cai, and K. K. Tamma. "Isogeometric generalized n th order perturbation-based stochastic method for exact geometric modeling of (composite) structures: Static and dynamic analysis with random material parameters." In: *Computer Methods in Applied Mechanics and Engineering* 346 (2019), pp. 1002–1024.

- **X. Hu**, J. Song, Z. Liao, Y. Liu, J. Gao, B. Menze, and W. Liu. "Morphological Residual Convolutional Neural Network (M-RCNN) for Intelligent Recognition of Wear Particles From Artificial Joints." In: *Friction* accepted (2021).

Peer-reviewed Conference Proceedings

- **X. Hu**, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu. "Face super-resolution guided by 3d facial priors." In: *European Conference on Computer Vision*. Springer. 2020, pp. 763–780.
- **X. Hu**, Y. Yan, W. Ren, H. Li, A. Bayat, Y. Zhao, and B. Menze. "Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features." In: *Medical Imaging with Deep Learning*. PMLR. 2021.
- Y. Zhao, Y. Liu, Y. Kan, A. Sekuboyina, D. Waldmannstetter, H. Li, **X. Hu**, X. Zhao, K. Shi, and B. Menze. "Spatial-Frequency Non-local Convolutional LSTM Network for pRCC Classification." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 22–30.
- Z. Zheng, W. Ren, X. Cao, **X. Hu**, T. Wang, F. Song, and X. Jia. "Ultra-High-Definition Image Dehazing via Multi-Guided Bilateral Learning." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, accepted.

Peer-reviewed Workshop Proceedings

- **X. Hu**, H. Li, Y. Zhao, C. Dong, B. H. Menze, and M. Piraud. "Hierarchical multi-class segmentation of glioma images using networks with multi-level activation function." In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 116–127.

Bibliography

[1] H. Brody. "Medical imaging." In: *Nature* 502.7473 (2013), S81–S81.

[2] A. Rajkomar, J. Dean, and I. Kohane. "Machine learning in medicine." In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.

[3] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.

[4] D. Shen, G. Wu, and H.-I. Suk. "Deep learning in medical image analysis." In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[5] A. Maier, C. Syben, T. Lasser, and C. Riess. "A gentle introduction to deep learning in medical image processing." In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 86–101.

[6] A. S. Lundervold and A. Lundervold. "An overview of deep learning in medical imaging focusing on MRI." In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 102–127.

[7] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás. "3D Deep Learning on Medical Images: A Review." In: *arXiv preprint arXiv:2004.00218* (2020).

[8] J. Ker, L. Wang, J. Rao, and T. Lim. "Deep learning applications in medical image analysis." In: *Ieee Access* 6 (2017), pp. 9375–9389.

[9] N. Dimitriou, O. Arandjelović, and P. D. Caie. "Deep learning for whole slide image analysis: An overview." In: *Frontiers in Medicine* 6 (2019).

[10] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, et al. "International evaluation of an AI system for breast cancer screening." In: *Nature* 577.7788 (2020), pp. 89–94.

[11] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." In: *Nature medicine* 25.6 (2019), pp. 954–961.

[12] L. Papp, C. P. Spielvogel, I. Rausch, M. Hacker, and T. Beyer. "Personalizing medicine through hybrid imaging and medical big data analysis." In: *Frontiers in Physics* 6 (2018), p. 51.

[13] A. Elnakib, G. Gimel'farb, J. S. Suri, and A. El-Baz. "Medical image segmentation: a brief survey." In: *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Springer, 2011, pp. 1–39.

[14] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[15] L. Wang, G. Li, F. Shi, X. Cao, C. Lian, D. Nie, M. Liu, H. Zhang, G. Li, Z. Wu, et al. "Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 411–419.

[16] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, et al. "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 415–423.

[17] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al. "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI." In: *Medical image analysis* 35 (2017), pp. 250–269.

[18] M. Bieth, L. Peter, S. G. Nekolla, M. Eiber, G. Langs, M. Schwaiger, and B. Menze. "Segmentation of skeleton and organs in whole-body CT images via iterative trilateration." In: *IEEE Transactions on Medical Imaging* 36.11 (2017), pp. 2276–2286.

[19] H. Li, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, and B. Menze. "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images." In: *NeuroImage* 183 (2018), pp. 650–665.

[20] M. Sezgin and B. Sankur. "Survey over image thresholding techniques and quantitative performance evaluation." In: *Journal of Electronic imaging* 13.1 (2004), pp. 146–166.

[21] S. Hojjatoleslami and F. Kruggel. "Segmentation of large brain lesions." In: *IEEE Transactions on Medical Imaging* 20.7 (2001), pp. 666–669.

[22] S.-Y. Wan and W. E. Higgins. "Symmetric region growing." In: *IEEE Transactions on Image processing* 12.9 (2003), pp. 1007–1015.

[23] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. "Multi-atlas segmentation with joint label fusion." In: *IEEE transactions on pattern analysis and machine intelligence* 35.3 (2013), pp. 611–623.

[24] E. Geremia, B. H. Menze, O. Clatz, E. Konukoglu, A. Criminisi, and N. Ayache. "Spatial decision forests for MS lesion segmentation in multi-channel MR images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 111–118.

[25] C.-C. Chang and C.-J. Lin. "LIBSVM: A library for support vector machines." In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.

[26] M. P. Heinrich and M. Blendowski. "Multi-organ segmentation using vantage point forests and binary context features." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 598–606.

[27] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, D. Rueckert, A. D. N. Initiative, et al. "Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling." In: *NeuroImage* 76 (2013), pp. 11–23.

[28] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[29] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." In: *nature* 521.7553 (2015), p. 436.

[30] D. Nie, L. Wang, Y. Gao, and D. Shen. "Fully convolutional networks for multi-modality isointense infant brain image segmentation." In: *2016 IEEE 13Th international symposium on biomedical imaging (ISBI)*. IEEE. 2016, pp. 1342–1345.

[31] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng. "3D deeply supervised network for automatic liver segmentation from CT volumes." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 149–157.

[32] L. Yu, X. Yang, H. Chen, J. Qin, P.-A. Heng, et al. "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3D MR images." In: *AAAI*. Vol. 17. 2017, pp. 36–72.

[33] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[34] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 424–432.

[35] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.

[36] G. Lin, A. Milan, C. Shen, and I. Reid. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1925–1934.

[37] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng. "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes." In: *IEEE transactions on medical imaging* 37.12 (2018), pp. 2663–2674.

[38] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis. "Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact Removal." In: *IEEE journal of biomedical and health informatics* 24.2 (2019), pp. 568–576.

[39] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation." In: *IEEE transactions on medical imaging* 39.6 (2019), pp. 1856–1867.

[40] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. "Efficient piecewise training of deep structured models for semantic segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3194–3203.

[41] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang. "Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation." In: *Neuroinformatics* 16.3 (2018), pp. 383–392.

[42] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li. "Spine-GAN: Semantic segmentation of multiple spinal structures." In: *Medical image analysis* 50 (2018), pp. 23–35.

[43] E. National Academies of Sciences, Medicine, et al. *Improving diagnosis in health care*. National Academies Press, 2015.

[44] B. Y. Reis, I. S. Kohane, and K. D. Mandl. "Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study." In: *Bmj* 339 (2009).

[45] A. Rajkomar, J. W. L. Yim, K. Grumbach, and A. Parekh. "Weighting primary care patient panel size: a novel electronic health record-derived measure using machine learning." In: *JMIR medical informatics* 4.4 (2016), e29.

[46] A. L. Beam and I. S. Kohane. "Big data and machine learning in health care." In: *Jama* 319.13 (2018), pp. 1317–1318.

[47] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al. "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy." In: *Radiology* 296.2 (2020), E65–E71.

[48] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." In: *nature* 542.7639 (2017), pp. 115–118.

[49] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." In: *Cell* 172.5 (2018), pp. 1122–1131.

[50] P. Blanc-Durand, L. Campedel, S. Mule, S. Jegou, A. Luciani, F. Pigneur, and E. Itti. "Prognostic value of anthropometric measures extracted from whole-body CT using deep learning in patients with non-small-cell lung cancer." In: *European radiology* (2020), pp. 1–10.

[51] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, and H. J. Aerts. "Deep learning predicts lung cancer treatment response from serial medical imaging." In: *Clinical Cancer Research* 25.11 (2019), pp. 3266–3275.

[52] Z. Tang, Y. Xu, L. Jin, A. Aibaidula, J. Lu, Z. Jiao, J. Wu, H. Zhang, and D. Shen. "Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients." In: *IEEE transactions on medical imaging* 39.6 (2020), pp. 2100–2109.

[53] H. Peng, D. Dong, M.-J. Fang, L. Li, L.-L. Tang, L. Chen, W.-F. Li, Y.-P. Mao, W. Fan, L.-Z. Liu, et al. "Prognostic value of deep learning PET/CT-based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma." In: *Clinical Cancer Research* 25.14 (2019), pp. 4271–4279.

[54] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[55] Y. Sui, O. Afacan, A. Gholipour, and S. K. Warfield. "Isotropic MRI super-resolution reconstruction with multi-scale gradient field prior." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 3–11.

[56] J. Lee, H. Kim, H. Chung, and J. C. Ye. "Deep learning fast MRI using channel attention in magnitude domain." In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 917–920.

[57] Y. Wu, Y. Ma, J. Liu, J. Du, and L. Xing. "Self-attention convolutional neural network for improved MR image reconstruction." In: *Information sciences* 490 (2019), pp. 317–328.

[58] C. Meurée, P. Maurel, J.-C. Ferré, and C. Barillot. "Patch-based super-resolution of arterial spin labeling magnetic resonance images." In: *Neuroimage* 189 (2019), pp. 85–94.

[59] F. Shi, J. Cheng, L. Wang, P.-T. Yap, and D. Shen. "LRTV: MR image super-resolution with low-rank and total variation regularizations." In: *IEEE transactions on medical imaging* 34.12 (2015), pp. 2459–2466.

[60] C. Dong, C. C. Loy, K. He, and X. Tang. "Learning a deep convolutional network for image super-resolution." In: *European conference on computer vision*. Springer. 2014, pp. 184–199.

[61] J. Shi, Z. Li, S. Ying, C. Wang, Q. Liu, Q. Zhang, and P. Yan. "MR image super-resolution via wide residual networks with fixed skip connection." In: *IEEE journal of biomedical and health informatics* 23.3 (2018), pp. 1129–1140.

[62] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander. "Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 611–619.

[63] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks." In: *Proceedings of the International Conference on Neural Information Processing Systems* (2014), pp. 2672–2680.

[64] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive growing of gans for improved quality, stability, and variation." In: *arXiv preprint arXiv:1710.10196* (2017).

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

[66] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8789–8797.

[67] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. "Improved techniques for training gans." In: *arXiv preprint arXiv:1606.03498* (2016).

[68] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. "Improved training of wasserstein GANs." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 5769–5779.

[69] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. "Photo-realistic single image super-resolution using a generative adversarial network." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[70] T. Kaneko, Y. Ushiku, and T. Harada. "Label-noise robust generative adversarial networks." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2467–2476.

[71] W. Hong, Z. Wang, M. Yang, and J. Yuan. "Conditional generative adversarial network for structured domain adaptation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1335–1344.

[72] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[73] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. "Medical image synthesis with deep convolutional adversarial networks." In: *IEEE Transactions on Biomedical Engineering* 65.12 (2018), pp. 2720–2730.

[74] K. Armanious, C. Jiang, S. Abdulatif, T. Küstner, S. Gatidis, and B. Yang. "Unsupervised medical image translation using Cycle-MedGAN." In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE. 2019, pp. 1–5.

[75] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur. "Image synthesis in multi-contrast MRI with conditional generative adversarial networks." In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2375–2388.

[76] A. Sharma and G. Hamarneh. "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network." In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1170–1183.

[77] M. E. Davis. "Glioblastoma: overview of disease and treatment." In: *Clinical journal of oncology nursing* 20.5 (2016), S2.

[78] F. Hanif, K. Muzaffar, K. Perveen, S. M. Malhi, and S. U. Simjee. "Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment." In: *Asian Pacific journal of cancer prevention: APJCP* 18.1 (2017), p. 3.

[79] A. Birbrair, A. Sattiraju, D. Zhu, G. Zulato, I. Batista, V. T. Nguyen, M. L. Messi, K. K. Solingapuram Sai, F. C. Marini, O. Delbono, et al. "Novel peripherally derived neural-like stem cells as therapeutic carriers for treating glioblastomas." In: *Stem cells translational medicine* 6.2 (2017), pp. 471–481.

[80] M. S. Nosrati and G. Hamarneh. "Local optimization based segmentation of spatially-recurring, multi-region objects with part configuration constraints." In: *IEEE transactions on medical imaging* 33.9 (2014), pp. 1845–1859.

[81] A. BenTaieb and G. Hamarneh. "Topology aware fully convolutional networks for histology gland segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 460–468.

[82] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren. "Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks." In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 64–76.

[83] S. Bauer, J. Tessier, O. Krieter, L.-P. Nolte, and M. Reyes. "Integrated spatio-temporal segmentation of longitudinal brain tumor imaging studies." In: *International MICCAI Workshop on Medical Computer Vision*. Springer. 2013, pp. 74–83.

[84] E. Alberts, G. Charpiat, Y. Tarabalka, T. Huber, M.-A. Weber, J. Bauer, C. Zimmer, and B. H. Menze. "A nonparametric growth model for brain tumor segmentation in longitudinal MR sequences." In: *BrainLes 2015*. Springer. 2015, pp. 69–79.

[85] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. "Deep learning markov random field for semantic segmentation." In: *IEEE transactions on pattern analysis and machine intelligence* 40.8 (2017), pp. 1814–1828.

[86] M. Piraud, A. Sekuboyina, and B. H. Menze. "Multi-level activation for segmentation of hierarchically-nested classes." In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018, pp. 0–0.

[87] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein. "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge." In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 287–297.

[88] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features." In: *Scientific data* 4.1 (2017), pp. 1–13.

[89] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge." In: *arXiv preprint arXiv:1811.02629* (2018).

[90] B. D. Cheson. "Role of functional imaging in the management of lymphoma." In: *Journal of Clinical Oncology* 29.14 (2011), pp. 1844–1854.

[91] W. K. Chan, W.-Y. Au, C.-Y. O. Wong, R. Liang, A. Y. Leung, Y.-L. Kwong, and P.-L. Khong. "Metabolic activity measured by F-18 FDG PET in natural killer-cell lymphoma compared to aggressive B-and T-cell lymphomas." In: *Clinical nuclear medicine* 35.8 (2010), pp. 571–575.

[92] P.-L. Khong, C. B. Pang, R. Liang, Y.-L. Kwong, and W.-Y. Au. "Fluorine-18 fluorodeoxyglucose positron emission tomography in mature T-cell and natural killer cell malignancies." In: *Annals of hematology* 87.8 (2008), pp. 613–621.

[93] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. "Brain tumor segmentation with deep neural networks." In: *Medical image analysis* 35 (2017), pp. 18–31.

[94] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." In: *Medical image analysis* 36 (2017), pp. 61–78.

[95] Z. Shi, G. Zeng, L. Zhang, X. Zhuang, L. Li, G. Yang, and G. Zheng. "Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 569–577.

[96] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[97] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[98] C.-C. Li, H.-F. Tien, J.-L. Tang, M. Yao, Y.-C. Chen, I.-J. Su, S.-M. Hsu, and R.-L. Hong. "Treatment outcome and pattern of failure in 77 patients with sinonasal natural killer/T-cell or T-cell lymphoma." In: *Cancer* 100.2 (2004), pp. 366–375.

[99] W.-Y. Au, S.-Y. Ma, C.-S. Chim, C. Choy, F. Loong, A. Lie, C. Lam, A. Leung, E. Tse, C.-C. Yau, et al. "Clinicopathologic features and treatment outcome of mature T-cell and natural killer-cell lymphomas diagnosed according to the World Health Organization classification scheme: a single center experience of 10 years." In: *Annals of Oncology* 16.2 (2005), pp. 206–214.

[100] C.-Y. Chen, M. Yao, J.-L. Tang, W. Tsay, C.-C. Wang, W.-C. Chou, I.-J. Su, F.-Y. Lee, M.-C. Liu, and H.-F. Tien. "Chromosomal abnormalities of 200 Chinese patients with non-Hodgkin's lymphoma in Taiwan: with special reference to T-cell lymphoma." In: *Annals of oncology* 15.7 (2004), pp. 1091–1096.

[101] S. H. Moon, S. K. Cho, W.-S. Kim, S. J. Kim, Y. C. Ahn, Y. S. Choe, K.-H. Lee, B.-T. Kim, and J. Y. Choi. "The role of 18F-FDG PET/CT for initial staging of nasal type natural killer/T-cell lymphoma: a comparison with conventional staging methods." In: *Journal of Nuclear Medicine* 54.7 (2013), pp. 1039–1044.

[102] X. Zhou, K. Lu, L. Geng, X. Li, Y. Jiang, and X. Wang. "Utility of PET/CT in the diagnosis and staging of extranodal natural killer/T-cell lymphoma: a systematic review and meta-analysis." In: *Medicine* 93.28 (2014).

[103] H. Fujiwara, Y. Maeda, Y. Nawa, M. Yamakura, D. Ennishi, Y. Miyazaki, K. Shinagawa, M. Hara, K. Matsue, and M. Tanimoto. "The utility of positron emission tomography/computed tomography in the staging of extranodal natural killer/T-cell lymphoma." In: *European journal of haematology* 87.2 (2011), pp. 123–129.

[104] H.-B. Wu, Q.-S. Wang, M.-F. Wang, H.-S. Li, W.-L. Zhou, X.-H. Ye, and Q.-Y. Wang. "Utility of 18F-FDG PET/CT for staging NK/T-cell lymphomas." In: *Nuclear medicine communications* 31.3 (2010), pp. 195–200.

[105] P.-L. Khong, B. Huang, E. Y. P. Lee, W. K. S. Chan, and Y.-L. Kwong. "Midtreatment 18F-FDG PET/CT scan for early response assessment of SMILE therapy in natural killer/T-cell lymphoma: a prospective study from a single center." In: *Journal of Nuclear Medicine* 55.6 (2014), pp. 911–916.

[106] R. Guo, P. Xu, H. Xu, Y. Miao, and B. Li. "The predictive value of pretreatment 18F-FDG PET/CT on treatment outcome in early-stage extranodal natural killer/T-cell lymphoma." In: *Leukemia & lymphoma* 61.11 (2020), pp. 2659–2664.

[107] B. Bai, H.-Q. Huang, Q.-C. Cai, W. Fan, X.-X. Wang, X. Zhang, Z.-X. Lin, Y. Gao, Y.-F. Xia, Y. Guo, et al. "Predictive value of pretreatment positron emission tomography/computed tomography in patients with newly diagnosed extranodal natural killer/T-cell lymphoma." In: *Medical oncology* 30.1 (2013), pp. 1–10.

[108] Y. Chang, X. Fu, Z. Sun, X. Xie, R. Wang, Z. Li, X. Zhang, G. Sheng, and M. Zhang. "Utility of baseline, interim and end-of-treatment 18 F-FDG PET/CT in extranodal natural killer/T-cell lymphoma patients treated with L-asparaginase/pegaspargase." In: *Scientific reports* 7.1 (2017), pp. 1–12.

[109] C. Jiang, X. Zhang, M. Jiang, L. Zou, M. Su, R. O. Kosik, and R. Tian. "Assessment of the prognostic capacity of pretreatment, interim, and post-therapy 18 F-FDG PET/CT in extranodal natural killer/T-cell lymphoma, nasal type." In: *Annals of nuclear medicine* 29.5 (2015), pp. 442–451.

[110] T. Sakai, M. C. Plessis, G. Niu, and M. Sugiyama. "Semi-supervised classification based on classification from positive and unlabeled data." In: *International conference on machine learning*. PMLR. 2017, pp. 2998–3006.

[111] M. Gori, G. Monfardini, and F. Scarselli. "A new model for learning in graph domains." In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* Vol. 2. IEEE. 2005, pp. 729–734.

[112] Y. Bai, H. Ding, S. Bian, T. Chen, Y. Sun, and W. Wang. "Simgnn: A neural network approach to fast graph similarity computation." In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 384–392.

[113] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. "Graph attention networks." In: *In: Proceedings of International Conference on Learning Representations (ICLR)* (2018).

[114] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. "A comprehensive survey on graph neural networks." In: *IEEE transactions on neural networks and learning systems* (2020).

[115] X. Hu, H. Li, Y. Zhao, C. Dong, B. H. Menze, and M. Piraud. "Hierarchical multi-class segmentation of glioma images using networks with multi-level activation function." In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 116–127.

[116] X. Hu, R. Guo, J. Chen, H. Li, D. Waldmannstetter, Y. Zhao, B. Li, K. Shi, and B. Menze. "Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images." In: *IEEE journal of biomedical and health informatics* 24.9 (2020), pp. 2599–2608.

[117] R. Guo*, **X, Hu***, H. Song*, P. Xu, H. Xu, A. Rominger, X. Lin, B. Menze, B. Li, and K. Shi. "Weakly supervised deep learning for determining the prognostic value of 18 F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type." In: *European journal of nuclear medicine and molecular imaging* (2021), pp. 1–11.

[118] X. Hu, Y. Yan, W. Ren, H. Li, A. Bayat, Y. Zhao, and B. Menze. "Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features." In: *Medical Imaging with Deep Learning*. PMLR. 2021.

[119] S. K. Zhou, H. Greenspan, and D. Shen. *Deep learning for medical image analysis*. Academic Press, 2017.

[120] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors." In: *nature* 323.6088 (1986), pp. 533–536.

[121] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[122] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[123] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).

[124] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[125] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. "Deformable convolutional networks." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 764–773.

[126] R. Zhang, F. Zhu, J. Liu, and G. Liu. "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis." In: *IEEE Transactions on Information Forensics and Security* 15 (2019), pp. 1138–1150.

[127] F. Yu and V. Koltun. "Multi-scale context aggregation by dilated convolutions." In: *arXiv preprint arXiv:1511.07122* (2015).

[128] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. "How does batch normalization help optimization?" In: 2018.

[129] Z. Zhang and M. R. Sabuncu. "Generalized cross entropy loss for training deep neural networks with noisy labels." In: 2018.

[130] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians." In: *Journal of the American Statistical Association* 112 (2017), pp. 859–877.

[131] X. Yi, E. Walia, and P. Babyn. "Generative adversarial network in medical imaging: A review." In: *Medical image analysis* 58 (2019), p. 101552.

[132] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks." In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[133] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for simplicity: The all convolutional net." In: *International Conference on Learning Representations workshop* (2014).

[134] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. "Explaining nonlinear classification decisions with deep taylor decomposition." In: *Pattern Recognition* 65 (2017), pp. 211–222.

[135] A. Kendall and Y. Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *International Conference on Neural Information Processing Systems* (2017).

[136] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. "Sanity checks for saliency maps." In: *Advances in Neural Information Processing Systems*. 2018, pp. 9505–9515.

[137] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze. "Evaluating saliency map explanations for convolutional neural networks: a user study." In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 2020, pp. 275–285.

[138] A. Borji. "Saliency prediction in the deep learning era: Successes and limitations." In: *IEEE transactions on pattern analysis and machine intelligence* (2019).

[139] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis." In: *Medical image analysis* 42 (2017), pp. 60–88.

[140] Y. Gou, B. Li, Z. Liu, S. Yang, and X. Peng. "CLEARER: Multi-Scale Neural Architecture Search for Image Restoration." In: *Advances in Neural Information Processing Systems* 33 (2020).

[141] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. "Progressive neural architecture search." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 19–34.

[142] H. Liu, K. Simonyan, and Y. Yang. "Darts: Differentiable architecture search." In: *In International Conference on Learning Representations* (2018).

[143] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le. "Regularized evolution for image classifier architecture search." In: *Proceedings of the aaai conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 4780–4789.

[144] H. Zhang, Y. Li, H. Chen, and C. Shen. "Memory-efficient hierarchical neural architecture search for image denoising." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3657–3666.

[145] Z. Liu, H. Wang, S. Zhang, G. Wang, and J. Qi. "NAS-SCAM: Neural Architecture Search-Based Spatial and Channel Joint Attention Module for Nuclei Semantic Segmentation and Classification." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 263–272.

[146] Y. Huang, X. Yang, R. Li, J. Qian, X. Huang, W. Shi, H. Dou, C. Chen, Y. Zhang, H. Luo, et al. "Searching Collaborative Agents for Multi-plane Localization in 3D Ultrasound." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 553–562.

[147] Y. Peng, L. Bi, M. Fulham, D. Feng, and J. Kim. "Multi-modality Information Fusion for Radiomics-Based Neural Architecture Search." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 763–771.

[148] X. Yan, W. Jiang, Y. Shi, and C. Zhuo. "Ms-nas: Multi-scale neural architecture search for medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 388–397.

[149] Z. Zhu, C. Liu, D. Yang, A. Yuille, and D. Xu. "V-nas: Neural architecture search for volumetric medical image segmentation." In: *2019 International Conference on 3D Vision (3DV)*. IEEE. 2019, pp. 240–248.

[150] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren. "Secure, privacy-preserving and federated machine learning in medical imaging." In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311.

[151] J. Konečn, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. "Federated learning: Strategies for improving communication efficiency." In: *arXiv preprint arXiv:1610.05492* (2016).

[152] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas. "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation." In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 92–104.

[153] Q. Chang, H. Qu, Y. Zhang, M. Sabuncu, C. Chen, T. Zhang, and D. N. Metaxas. "Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13856–13866.