UNIVERSITY OF CALIFORNIA
Los Angeles

Improving Acute Ischemic Stroke Diagnosis Using Medical Imaging and Deep Learning Methods

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Haoyue Zhang

2023

ABSTRACT OF THE DISSERTATION

Improving Acute Ischemic Stroke Diagnosis Using Medical Imaging and Deep Learning Methods

by

Haoyue Zhang
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2023
Professor Corey W. Arnold, Chair

Acute ischemic stroke (AIS) is a cerebrovascular disease caused by deceased blood flow in the brain. Treatment of AIS is heavily dependent on the time since stroke onset (TSS), either by clock time or tissue time. AIS treatments aim to restore blood flow in the stroke-affected area to minimize infarction. Current clinical guidelines recommend thrombolytic therapies (e.g. Intravenous(IV) or Intra-arterial (IA) tissue Plasminogen Activator (tPA) for patients presenting within 4.5 hours of TSS and Mechanical Thrombectomy (MTB) (e.g. surgical removal of the clot) for patients with TSS up to 24 hours. This research attempts to use both CT and MRI to predict the eligibility of AIS patients and their response to treatment while addressing several challenges in neuroimaging and AIS diagnosis in clinical settings using novel machine learning and deep learning approaches. A Self-supervised Learning approach, called intra-domain task-adaptive transfer learning, is the first proposed to predict TSS using limited training data. A hybrid transformer model that utilizes spatial neighborhood information in brain regions is proposed to predict MTB success. A pure transformer and a specifically designed Masked Image Model are developed to predict Large Vessel Occlu-

sion (LVO). Last, a transformer-based super-resolution framework is proposed to generate synthesized thin-slice images from thick-slice images. Together, these models demonstrate the effectiveness of the attention mechanism and the usefulness of self-supervised learning for clinical deep learning applications given the limited data resources compared to natural images.

The dissertation of Haoyue Zhang is approved.

Yingnian Wu

Holden H. Wu

Dan Ruan

William F. Speier

Corey W. Arnold, Committee Chair

University of California, Los Angeles

2023

Dedicated to my family and friends

Special gratitude to my beloved parents, my wife, and my children

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my utmost gratitude to my Ph.D. advisor, Dr. Corey Arnold for his continuous support and guidance throughout my Ph.D. study. I would not have switched to research and academia without his inspiration. I also want to thank my other committee members: Dr. William Speier, Dr. Dan Ruan, Dr. Yingnian Wu, and Dr. Holden Wu for their expertise from fundamentals to applications, for their kindness and enthusiasm, and for their patience and understanding. I want to thank Dr. Kambiz Nael for providing constructive suggestions and clinical insights for the projects. I am extremely thankful to my colleague, Dr. Jennifer Polson for her help, clinically and technically. Her dedication and focus inspired me during my Ph.D. journey. I would also like to acknowledge all current and past members of the computational diagnostics (CDx) group and medical imaging informatics (MII) group for their support. Special thanks to my colleagues Zichen Wang, Wenyuan Li, Jiayun Li, Karthik Sarma, Johnny Ho, Yimen Meng, Shiwen Shen, Lew Andrada, Yanan Lin, Leihao Wei, Tianran Zhang, Carlos Olivares, Saarang Panchavati, Mara Pleasure, Ashwath Radhachandran, Katya Redekop, Nathan Siu, Eric Yang, Ilyass Majji, Simon Han, David Gordon, Nova Smedley, Panayiotis Petousis, and many more. Special thanks to our staff members Shawn Chen, Isabel Rippy, Denise Luna, and Rushi Kulkarni for all their help and support. In addition, I want to thank our collaborators Dr. John Hoffman, Dr. Suzie El-Saden, Dr. Noriko Salamon, Dr. Bryan Yoo, Dr. Shingo Kihira, and Dr. Iris Chen.

I would also like to thank the UCLA Graduate Division, UCLA Bioengineering Department, and NIH for providing financial support through my Ph.D. study. This research is mainly supported by NINDS R01NS100806 and Dissertation Year Fellowship.

Last, I would like to thank my friends and my family. They are the most invaluable assets in my life and no words can easily express my gratitude.

VITA

2010–2014	B.S. (Mathematics/Economics, Statistics), UCLA
2014-2016	M.S. (Computational Science and Engineering), Rice University
2016-2017	Data Visualization Engineer, Acumen LLC
2017-2018	Data Scientist, UCLA Health - Radiology
2018-2023	PhD. Bioengineering (expected), UCLA
2021-2023	Machine Learning Intern and research collaborator, Infervision
2022-2023	Machine Learning Intern, Genentech/Roche

SELECTED PUBLICATIONS

[1] Yang, X., Yu, P., **Zhang, H.**, Zhang, R., Liu, Y., Li, H., ... & Yang, Q. (2023). Deep Learning Algorithm Enables Cerebral Venous Thrombosis Detection With Routine Brain Magnetic Resonance Imaging. Stroke, 54(5), 1357-1366.

Polson, J. S.*, **Zhang, H.***, Nael, K., Salamon, N., Yoo, B. Y., El-Saden, S., ... & Arnold, C. W. (2022). Identifying acute ischemic stroke patients within the thrombolytic treatment window using deep learning. Journal of Neuroimaging, 32(6), 1153-1160.

[1]*denotes equal contribution

Yu, P.*, **Zhang, H.***, Kang, H., Tang, W., Arnold, C. W., & Zhang, R. (2022, September). RPLHR-CT Dataset and Transformer Baseline for Volumetric Super-Resolution from CT Scans. MICCAI 2022 Proceedings, Part VI (pp. 344-353)

Tang, W.*, **Zhang, H.***,... & Zhang, R. (2022, July). MMMNA-Net for Overall Survival Time Prediction of Brain Tumor Patients. International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)

Zhang, H.*, Polson, J. S.*, ... & Arnold, C. W.. "Predicting Thrombectomy Recanalization from CT Imaging Using Deep Learning Models." 2022 Medical Imaging with Deep Learning (2022).

Polson, J.*, **Zhang, H.***, ... & Arnold, C. W. (2021, November). A Semi-Supervised Learning Framework to Leverage Proxy Information for Stroke MRI Analysis. International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)

Zhang, H.*, Polson, J.*, ... & Arnold, C. (2021, July). A machine learning approach to predict acute ischemic stroke thrombectomy reperfusion using discriminative mr image features. IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)

Zhang, H.*, Polson, J. S.*, ... & Arnold, C. W. (2021). Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time. Computerized Medical Imaging and Graphics, 90, 101926.

Ho, K. C., Speier, W., **Zhang, H.***, Scalzo, F., El-Saden, S., & Arnold, C. W. (2019). A machine learning approach for classifying ischemic stroke onset time from imaging. IEEE transactions on medical imaging, 38(7), 1666-1676.

CHAPTER 1

Introduction

Stroke is a cerebrovascular disease accounting for 2.7 million deaths worldwide every year. In the United States, it is the fifth leading cause of death, with approximately 795,000 people having a stroke and out of which 140,000 cases lead to death per year [Ben19]. Patients who have experienced strokes can suffer from severe and life-altering consequences. The physical and cognitive impairments that can result from a stroke, such as paralysis, communication difficulties, and cognitive impairment, can have long-lasting and detrimental effects on the patient's quality of life and ability to live independently. These disabilities can also lead to institutionalization and long-term care, which can pose a significant economic burden on healthcare systems and society as a whole. Therefore, there is a strong incentive to minimize the incidence of stroke and improve patient outcomes, given its high frequency, impact, and cost. There are two types of stroke: ischemic stroke and hemorrhage stroke, where 87% are ischemic strokes. For Acute Ischemic Stroke (AIS), timely and accurate diagnosis and treatment are crucial for a better outcome. Treatment of AIS is heavily dependent on the time since stroke onset (TSS), either by clock time or tissue time [RT14]. AIS treatments aim to restore blood flow in the stroke-affected area to minimize infarction. Current clinical guidelines recommend thrombolytic therapies (e.g. Intravenous(IV) or Intra-arterial (IA) tissue plasminogen activator (tPA) for patients presenting within 4.5 hours of TSS and mechanical thrombectomy (e.g. surgical removal of a clot) for patients with TSS up to 24 hours [TFO14, UFZ18].

1.1 Motivation

1.1.1 Identifying Patients Within Thrombolytic Treatment Window

The goal of thrombolysis as a medical intervention is to dissolve blood clots that have developed within blood vessels. The current clinical guidelines recommend it be used on patients with a known symptom onset time (time since stroke TSS) within 4.5 hours [PRA19, DKA16]. Favorable outcome is expected for patients within the treatment window [LBK10]. On the other hand, administration of thrombolysis outside of the treatment window may lead to hemorrhage or even worse, mortality. More than 35% of the AIS patients have an unknown TSS [UFZ18] and only 6.5% of patients hospitalized for AIS in the United States have received intravenous thrombolysis, and unknown TSS has been determined as the primary reason for treatment exclusion [JWG16]. Extensive research projects and clinical trials have been conducted to extend the eligibility of AIS patients for thrombolysis, many of them are using imaging, such as DWI-FLAIR mismatch, Diffusion-perfusion mismatch, etc [TSB18, NJH18, AMK18, BZL19].

Within minutes of a stroke, diffusion-weighted imaging (DWI) can identify reduced apparent diffusion coefficient (ADC) of ischemic lesions, while fluid-attenuated inversion recovery (FLAIR) can reveal a net increase in water contents within 1 to 4 hours. This difference in signals between DWI and FLAIR can be utilized as a tissue clock. However, this tissue clock signal does not necessarily provide reliable TSS information and the inter-reader agreement is moderate to low. Given this situation, Machine Learning (ML) has shown great potential in predicting TSS using MRI, either using traditional radiomics features and ML models [LLH20] or using end-to-end approaches with Deep Learning (DL) models [HSZ19, ZJZ21]. However, previous work either requires infarct core segmentation or the use of perfusion images, limiting the usage in real-world clinical settings.

1.1.2 Predicting Thrombectomy Success Using Pre-treatment Imaging

Endovascular thrombectomy (EVT) is a medical procedure used to remove blood clots from blocked blood vessels in the brain or other parts of the body, typically by inserting a catheter through an artery into the site of the clot and restoring the blood flow of the infarct area. Successful EVT has many potential factors that could influence a patient's response to treatment. In practice, success is measured by restoration of blood flow to the stroke area, quantified by the modified treatment in cerebral infarction (mTICI) score [NJH18, AMK18, Ban11]. Clinical trials and other studies have illustrated that patients who experience partial and/or full recanalization of the blood vessel typically experience better outcomes, particularly if recanalization is achieved in three attempts or less [DCB17, GTF19, VCS17]. Pre-treatment MRI and CT may provide useful information regarding the success of endovascular thrombectomy, providing critical information for neurosurgeons before they create a treatment plan for AIS patients.

1.1.3 Predicting Large Vessel Occlusion

Large Vessel Occlusion (LVO) stroke is a type of ischemic stroke with complete or partial blockage of a major blood vessel in the brain. The occlusion can disrupt blood flow to a significant area of the brain, leading to a stroke or other serious neurological complications. LVO accounts for 24% to 46% of the AIS [RWS19] and EVT has shown to be far more effective than thrombolysis for LVO strokes [BFB15, CMK15, GDM15, JCC15]. Timely determination of LVO stroke is crucial for the following EVT treatment and outcome [MHO20].

1.2 Challenges

Deep Learning (DL) has demonstrated its superiority in both computer vision and natural language processing tasks. Particularly, in medical imaging, CNN-based and recently

proposed transformer-based architectures are widely used in different modalities, organs, diagnosis, detection, and lesion segmentation. However, DL approaches require a large amount of training data. Depending on the specific tasks, the data required to let the model converge can scale up easily. Particularly, the recent success of the vision transformer is established on an even larger data scale. On the other hand, medical images, particularly stroke-related neuroimaging, are scarce due to the limitation and regulations of medical institutions, rare diseases, and different protocols. Therefore, the first challenge is that we need to effectively train DL models with limited data. The second challenge is the significant human effort required for annotation, especially for tasks involving segmentation and detection. To achieve optimal performance, certain disease diagnosis, classification, or prediction tasks may require lesion segmentation to enable the model to concentrate on the specific area of interest rather than the entire input. The third hurdle pertains specifically to AIS and involves the need for the algorithm to produce results in a timely manner within the fast-paced environment of real-world clinical settings. The utilization of AI algorithms may be restricted due to the low quality of images and the thick slice thickness associated with the AIS imaging protocol, which presents the fourth challenge.

1.3 Contributions

The main contributions of this dissertation can be summarized in the following specific aims:

Aim 1. Develop an intra-domain task-specific self-supervised learning approach and attention based 2D and 3D CNN models to classify time since stroke using diffusion weighted MRI.

a. To Investigate the effectiveness of different pretraining approaches, including self-supervised and supervised pretraining tasks to help the model converge on a relatively small dataset.

b. To develop attention-based 2D and 3D CNN models that achieve end-to-end

training to predict TSS and evaluate on an external center.

c. To propose a semi-supervised approach to predict DWI-FLAIR mismatch from the TSS model.

Aim 2. Develop a CNN-transformer hybrid model to predict EVT outcome using both CT and MRI.

a. To develop radiomics-based ML models to examine the feasibility of predicting mTICI using pre-treatment MRI.

b. To develop a CNN-transformer hybrid model to predict mTICI using non-contrast CT and CT angiography.

c. To further add contrastive self-supervised learning pretraining to the model and evaluate the performance on both CT and MRI.

Aim 3. Develop a pure vision transformer model to predict large vessel occlusion.

a. To develop a swin transformer model to predict large vessel occlusion using non-contrast CT.

b. To develop a masked imaging self-supervised learning approach for pretraining the model.

Aim 4. Develop a transformer-based super-resolution model to synthesize 1mm slice thickness from 5mm slice thickness.

a. To develop a transformer-based super-resolution model to generate 1mm slice from 5mm slice CT images using real-paired CT data.

b. To further improve the model with better performance and efficiency and evaluate on multiple external centers.

1.4 Outline of the Dissertation

This dissertation is organized as follows:

Chapter 2 provides the background on acute ischemic stroke, the diagnosis, and treatment, the role of medical imaging in stroke, ML and DL, attention mechanism, transformers, and a selected review of related work.

Chapter 3 presents a novel intra-domain task-adaptive self-supervised approach to predict time since stroke using MRI, composed of previously published work.

Chapter 4 summarizes the efforts to predict thrombectomy outcomes using ML and DL approaches, including published work.

Chapter 5 demonstrates a transformer-based self-supervised learning approach to predict large vessel occlusion from non-contrast CT data.

Chapter 6 builds a transformer-based super-resolution algorithm to generate thin-sliced CT images from thick-sliced CT images.

Chapter 7 concludes the dissertation and discusses limitations and potential future directions.

CHAPTER 2

Background

2.1 Acute Ischemic Stroke

Stroke is the second leading cause of death and the second largest healthcare burden estimated in disability-adjusted life-years globally. In terms of stroke incidence, the highest rates were observed in East Asia, followed by the Eastern European region, while the lowest rates were reported in central Latin America. The age-specific incidence rates were similar between women and men up to the age of 55 years, but men had higher rates than women between 55 and 75 years of age, with the rates leveling out at older ages [Gor19]. 87% of the strokes are ischemic strokes and 13% are hemorrhage strokes. Figure 2.1 illustrate the two types of stroke. This dissertation focuses on acute ischemic stroke (AIS), a condition in which one or multiple occlusions narrow or block arteries to the brain, leading to severely reduced blood flow (ischemia) and tissue death. On the other hand, hemorrhagic stroke occurs when a weakened blood vessel ruptures, causing bleeding inside the brain. When AIS happens, the clot blocked the artery from supplying blood to the brain, causing brain cells to become deprived of oxygen and nutrients, leading to dysfunction and eventual death if the blood flow is not restored quickly. The symptoms of acute ischemic stroke can vary depending on the location and extent of the blockage, but they may include sudden weakness, numbness, or paralysis in the face, arm, or leg, usually on one side of the body, difficulty speaking or understanding speech, vision problems, severe headaches, and loss of balance or coordination [Wal22]. The infarct core is the region of the stroke where the brain tissues are already infarcted and irreversibly dead regardless of reperfusion. The penumbra area

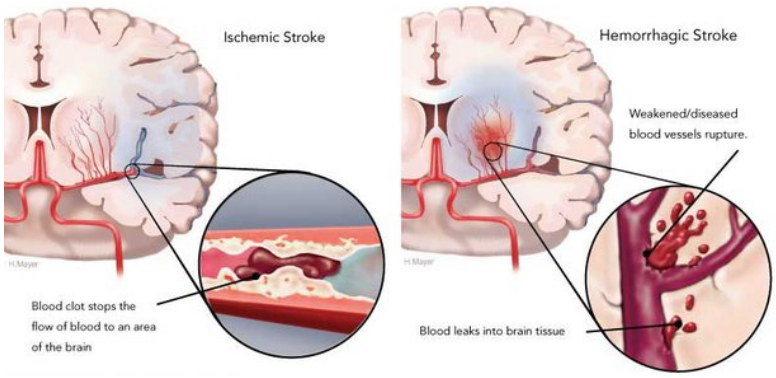


Figure 2.1: Two types of stroke. This figure is credited to [Jaf19]

is defined as the regions where the brain tissues are at risk of becoming infarcted. There are mainly two types of treatment, thrombolysis, and thrombectomy. The treatment goal is to remove the clot and restore the blood flow in a timely manner so that the extent of the damage can be minimized. The specific treatment options depend on various factors, such as the severity and location of the stroke and the time since the onset of symptoms. The design of the treatment plan follows the guidelines from American Heart Association/American Stroke Association [PRA19].

2.1.1 Thrombolysis

Thrombolysis is a medical treatment that involves the use of drugs to dissolve blood clots that have formed inside blood vessels. This treatment is commonly used for acute ischemic stroke and other conditions caused by blood clots, such as deep vein thrombosis and pulmonary embolism.

The most common thrombolytic drug used for acute ischemic stroke is Intravenous (IV) or Intra-arterial (IA) tissue plasminogen activator (tPA). tPA is a protein that occurs naturally in the body and helps to break down blood clots. When administered as a medication, tPA can dissolve the clot causing the stroke and restore blood flow to the affected area

of the brain. Thrombolysis may be administered through an intravenous infusion or by directly injecting the medication into the affected artery. This treatment requires close monitoring in a hospital setting and may be followed by other treatments, such as mechanical thrombectomy or supportive care.

Thrombolysis is most effective when administered as soon as possible after the onset of symptoms (or Time since stroke, TSS) of acute ischemic stroke. This is because the longer the blood flow is blocked, the greater the risk of irreversible brain damage. However, thrombolysis also carries a risk of bleeding, so it must be carefully considered in each individual case. Efforts have been made to expand the treatment window of thrombolysis to 4.5 hours [HKB08] and onset within 4.5 hours is currently used in clinical guidelines. Multiple pieces of research aimed to extend the treatment window to up to 6 to 9 hours [Ahm13, RSS02, MCP19] with the guidance of imaging. Due to the trade-off between the risk of hemorrhage for AIS patients with longer onset time and benefits directly related to outcome, tPA administration requires reliable symptom onset time. However, up to 35% of the AIS population have unknown onset time [EBS18]. Per clinical Guidelines, patients with unknown TSS are excluded from thrombolysis treatment. Two recent clinical trials have provided solutions for unknown TSS using MRI. The WAKE-UP trial uses DWI-FLAIR mismatch to differentiate the signals in DWI and FLAIR sequences to determine a patient's eligibility for thrombolysis [Tho11, TFO14, TSB18]. The WAKE-UP trial showed that AIS patients with unknown TSS that were treated by tPA achieved significantly better functional outcomes. MR WITNESS trial uses a quantitative mismatch of DWI and FLAIR (qDFM) to treat AIS patients with a median TSS of 11.2 hours using thrombolytics and demonstrated the safety of the treatment beyond recommended time windows [SWS18]. Therefore, using MRI to extend the eligibility of thrombolysis is practical in a real-world setting.

2.1.2 Thrombectomy

Thrombectomy, also known as endovascular or mechanical thrombectomy (EVT or MTB), is a minimally invasive medical procedure in which a blood clot, also known as a thrombus, is removed from a blood vessel in the brain. This procedure is typically performed in patients who have had an ischemic stroke, which is caused by a blockage in a blood vessel in the brain. Before the procedure, the patient is typically given a sedative to help them relax. The patient's vital signs, such as blood pressure and heart rate, are monitored throughout the procedure. The physician may also administer an anticoagulant medication to help prevent future blood clots. The physician makes a small incision in the patient's groin to access the femoral artery. A thin, flexible catheter is then inserted through the incision and guided up to the site of the clot in the brain. Using real-time imaging, such as Digital Subtraction Angiography (DSA) [KBW19], the physician is able to locate the clot and determine the best approach for removing it. There are several types of mechanical devices that can be used to remove the clot, including stent retrievers and aspiration catheters [FS16]. Stent retrievers are mesh-like devices that are threaded through the catheter and positioned around the clot. Once in place, the retriever is expanded, trapping the clot within its mesh. The physician then gently pulls the retriever and clot back through the catheter and out of the body. Aspiration catheters work by using suction to draw the clot into a catheter and out of the body. Once the clot has been removed, the physician monitors the patient's vital signs and neurological status to ensure that there are no complications. Once the procedure is complete, the catheter is removed and pressure is applied to the incision site to stop any bleeding. The patient may be required to lie flat for a period of time to prevent bleeding.

EVT is most effective when performed as quickly as possible after the onset of stroke symptoms, ideally within six hours of symptom onset. However, in some cases, it may still be beneficial to perform an EVT up to 24 hours after the onset of symptoms [AMK18, Ben19]. Specifically, the DAWN trial showed that AIS patient whose TSS was between 6 to 24 hours and had a mismatch between clinical deficit and infarct achieved better outcomes through

EVT than standard care [NJH18]. Diffusion-weighted MRI or perfusion CT was used for the measurement of clinical deficit and infarct volume. Further study showed that advanced imaging such as perfusion CT does not provide extra benefit with regard to clinical outcomes, deeming non-contrast CT and CT angiography enough for the assessment [NHL21]. On the other hand, DEFUSE 3 trial uses CT perfusion or MR diffusion perfusion imaging to assess EVT eligibility and showed that patients with TSS of 6 to 16 hours had better outcomes following EVT plus standard care [AMK18]. and EVT is considered to be a highly effective treatment for ischemic stroke, particularly for patients who are not able to receive intravenous (IV) thrombolytic therapy, or who have not responded to IV thrombolytic therapy [LBG17].

The success of EVT is evaluated by Thrombolysis in Cerebral Infarction (TICI) score, proposed by [HF03], modified from the Thrombolysis in Myocardial Infarction (TIMI) scale. TIMI is a scoring system used to assess the severity of coronary artery disease and predict the risk of adverse outcomes in patients with acute coronary syndromes (ACS), including myocardial infarction (MI). TICI score evaluates the extent of blood flow restoration in the brain after each attempt of EVT using Angiography. The scale range from no perfusion (grade 0) to complete perfusion (grade 3). modified TICI (mTICI) score was proposed to amend the grading system to better reflect different grades of recanalization. Table 2.1 showed the details of the mTICI score. During EVT, if a complete or near-complete recanalization is achieved after one attempt of clot retrieval, it is called First-pass Effect (FPE). Research has shown that patients who experienced FPE correlated with significantly improved clinical outcomes, decreased mortality, and a significantly lower rate of hemorrhagic transformation [ZCL18, JCW19, DPG20].

2.1.3 Large Vessel Occlusion

Large vessel occlusion (LVO) refers to the complete or partial blockage of a major blood vessel in the brain, such as the middle cerebral artery (MCA), internal carotid artery (ICA), or basilar artery. LVO is a medical emergency that requires immediate attention and treatment,

Table 2.1: Modified TICI score

Score	Modified Thrombolysis in Cerebral Infarction Scale
0	No perfusion
1	Penetration but no perfusion. Antegrade reperfusion past the initial occlusion, but limited distal branch filling with little or slow distal reperfusion
2a	Antegrade reperfusion of less than half of the occluded target artery previously ischemic territory (e.g. in 1 major division of the MCA and its territory)
2b	Antegrade reperfusion of more than half of the previously occluded target artery ischemic territory (e.g. in 2 major divisions of the MCA and their territories)
2c	Near-complete perfusion except for slow flow in a few distal cortical vessels or presence of small distal cortical emboli
3	Complete antegrade reperfusion of the previously occluded target artery ischemic territory, with an absence of visualized occlusion in all distal branches.

as it can lead to severe brain damage or even death.

The most common cause of LVO is a blood clot that forms inside a blood vessel in the brain. This can occur due to a variety of conditions, including atherosclerosis (narrowing and hardening of the arteries), embolism (the blockage of a blood vessel by a foreign object such as a blood clot or air bubble), or thrombosis (the formation of a blood clot within a blood vessel) [VAB20, HTC13]. As shown in figure 2.2, up to 46% of AIS are LVO.

The symptoms of LVO can vary depending on the location and severity of the blockage. Common symptoms may include sudden weakness or numbness on contralateral hemibody and face, contralateral homonymous hemianopsia, ipsilateral gaze deviation, aphasia, neglect, dizziness, nausea, vomiting, gait, and balance issues, etc [RWS19].

If LVO is suspected, immediate medical attention is necessary. EVT is the standard

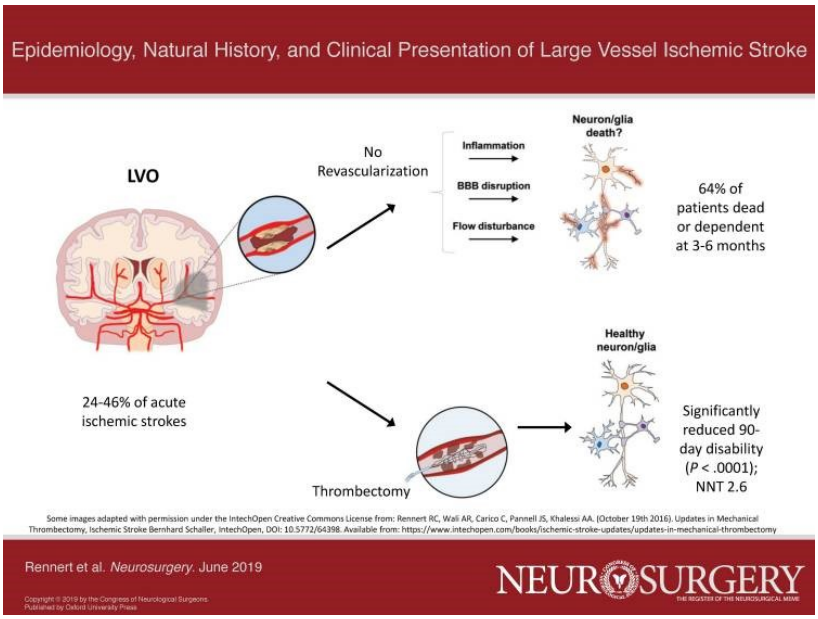


Figure 2.2: A brief illustration of large vessel occlusion. This figure is credited to [RWS19]

treatment for stroke with LVO [GTF19]. In fact, LVO is a major inclusion criterion for EVT treatment. Overall, LVO is a serious condition that requires urgent medical attention to prevent potentially life-threatening complications. Early detection and treatment can greatly improve the chances of a positive outcome and minimize the risk of long-term disability. EVT has been shown to significantly improve outcomes for patients with LVO when performed within a certain time window. However, it is not appropriate for all types of stroke or all patients. It is important to determine if a stroke patient has large vessel occlusion (LVO) because it can affect the course of treatment and the patient's prognosis. LVO is associated with more severe strokes and can result in a higher risk of disability or death and early identification of LVO can help guide the choice of treatment. In addition, identifying LVO can also help in the decision-making process for other treatments, such as thrombolytic drugs or antiplatelet medications. These treatments can be effective in certain cases but may increase the risk of bleeding in patients with LVO, and careful consideration of the risks and benefits is necessary.

Currently, there are different approaches to determining LVO before treatment [NIM22]. The gold standard to confirm the presence of LVO is digital subtraction angiography (DSA). The Los Angeles Motor Scale (LAMS) is used in the pre-hospital environment and showed good sensitivity (76%) and low specificity (65%) for a cut-off of ¿= 4 [NSS18]. The Cincinnati Prehospital Stroke Scale (CPSS) is another pre-hospital scoring system that showed a better specificity (88%) but lower sensitivity (41%) for a cut-off score of 3 [RHT18]. The Rapid Arterial oCclusion Evaluation (RACE) scale is a recently developed scoring system with a sensitivity of 85% and specificity of 73% [CCC17, LSD20]. Glasgow Coma Scale (GCS) has shown very high sensitivity (94%) and specificity (90%) for a cut-off of ¡ 15 [HHJ21].NIH Score Scale (NIHSS) is also modified and adopted for LVO prediction in multiple studies [PHR17, HHB16]. There are also several novel scales that were developed recently but further investigation is needed to verify their robustness [TVC17, TSS19, VAF19]. Other than a scoring system, imaging and physiological monitoring methods are also commonly used. Computed tomography angiography (CTA) is most widely used in many centers with very high sensitivity (83% to 97%) and specificity (87% to 99%) [LLR21, BJB20, FBH21]. On the MRI side, black-blood MRI is a unique modality that demonstrated high diagnostic accuracy and reliability with 100% sensitivity and specificity [AAE19]. Fluid Attenuated Inversion Recovery (FLAIR) also showed good sensitivity (98%) and specificity (86%) [BTL20].

2.2 Stroke Imaging

Radiology plays a critical role in stroke diagnosis and treatment. CT and MRI are the most commonly used in clinical settings. MRI provides a more sensitive AIS diagnosis than CT while non-contrast CT (NCCT) is good at quickly ruling out hemorrhage at the initial stage during admission. Studies show that standard MRI is five times more sensitive and twice more accurate than NCCT in diagnosing AIS and both MRI and NCCT perform similarly at diagnosing hemorrhage [CKN07]. However, historically more centers use CT machines for

standard stroke diagnosis. Cost, time, and availability are several major factors that lead to more usage of CT than MRI. [PSL19] showed that although MRI scan duration is longer than CT, the overall time frame for imaging diagnostics of AIS patients with MRI is comparable to CT, not delaying the treatment or impacting the outcomes. Typical MRI series for stroke protocol include Diffusion imaging such as Diffusion Weighted Imaging (DWI) at B0 and B1000, Apparant Diffusion Coefficient (ADC), Fluid-attenuated Inversion Recovery (FLAIR), Perfusion MRI such as raw Perfusion Weighted Imaging (PWI) and perfusion parameter maps including Cerebral Blood Flow (CBF), Cerebral Blood Flow (CBV), Mean Transit Time (MTT), Time-to-peak (TTP) and Time-to-Maximum (Tmax). Typical CT series for stroke protocol include NCCT, CT Angiography (CTA), and CT perfusion with the same parameter maps as in perfusion MRI.

2.2.1 Magnetic Resonance Imaging

MRI (Magnetic Resonance Imaging) is a noninvasive medical imaging technique that uses strong magnetic fields and radio waves to generate detailed images of internal body structures. The fundamentals of MRI are based on the physical properties of the body's atomic nuclei specifically, their interaction with magnetic fields [PK12].

When a patient is placed in a strong magnetic field, the hydrogen atoms in their body align with the magnetic field, producing a net magnetic moment. By applying a brief pulse of Radiofrequency (RF) energy, these hydrogen atoms can be excited and emit their own radio frequency signals, which are detected by the MRI scanner. The nuclei return to the resting alignment through relaxation. After a while, the emitted signals are measured. These signals provide information about the location and intensity of the magnetic field in the body, which can be used to generate images through Fourier transformation. Different types of tissue generate different signals, depending on their chemical composition and physical properties. By analyzing these signals, MRI can provide detailed images of soft tissues, such as the brain, spinal cord, and internal organs, as well as bones and other hard tissues. By varying

the RF pulse sequences and measurements, MRI can be used to selectively encode spatial information into the signals generated by the hydrogen atoms to highlight different types of tissues. This allows for the generation of detailed, three-dimensional images of internal structures.

T1-weighted scans and T2-weighted scans are the two most common MRI sequences. Short TE and TR times are used to produce T1-weighted images. On the other hand, longer TE and TR times are utilized to produce T2-weighted images. The contrast and brightness of different tissues displayed differently on T1-weighted and T2-weighted sequences. For example, Cerebrospinal Fluid (CSF) is dark on the T1-weighted sequence and bright on the T2-weighted sequence.

2.2.1.1 Diffusion Weighted Imaging

Diffusion Weighted Imaging (DWI) is designed to detect the Brownian (random) movement of water molecules within a voxel of tissue. Due to the influence of different cell structures, water molecules cannot move freely. Inside the human body, water in the extracellular environment can move relatively freely or is called diffused freely while diffusing restrictively in the intracellular environment. During ischemia of tissues, water moves from the extracellular to the intracellular environment due to the osmotic gradient and when this happens, the water movement is restricted intracellularly, therefore showing a bright signal on DWI by measuring the attenuation of the T2 signal based on how easily water molecules are able to diffuse in that region. The more easily water can diffuse, the less initial T2 signal will remain. Simply put, DWI is generated by applying a diffusion-related gradient pulse to a standard MRI sequence. Therefore, DWI is very sensitive in detecting infarct of AIS, within minutes. [BDS16]. Clinically for stroke diagnoses, DWI is widely used for early identification of ischemic stroke and differentiation of acute from chronic stroke. The DWI generation is as follows: First, obtain a T2-weighted image with no diffusion attenuation (b=0). Next, the ease with which water can diffuse is assessed in various directions (x,y,z). Next, b-value

generation by applying a strong gradient symmetrically on either side of the 180-degree pulse. The b-value measures the degree of diffusion weighting applied, the higher the b-value, the stronger the signal attenuation. Eventually, four sets of images are generated: T2 B=0 images and three DWI images in x, y, and z directions with the T2 signal attenuated according to how easily the water diffuses in that direction [KP06]. For stroke, B=500 and B=1000 sequences are usually used. The Apparent Diffusion Coefficient (ADC) Map can be generated by the log of isotropic DWI divided by the initial T2 signal:

$$ADC = -ln\frac{S/S_0}{b} \tag{2.1}$$

where S is the signal intensity at the given b value, S_0 is the intensity of no diffusion gradient and b is the b value. Commonly, DWI and ADC are used together by radiologists to identify infarct core where it is bright on DWI and dark on ADC [Lai14].

2.2.1.2 Fluid Attenuated Inversion Recovery

Introduced by Hajnal et al in 1992 for brain diagnostics [HBK92], Fluid Attenuated Inversion Recovery (FLAIR) is a special inversion recovery (IR) series with a long inversion time (TI) before the signal is acquired, followed by a short TI to null the signal from CSF. Therefore, CSF is shown as dark while preserving signals from other brain tissues thus FLAIR remains similar to T2-weighted images with CSF inverted. Briefly, we can interpret FLAIR as T2-weighted images with free-flowing water as dark and non-free-flowing water and fat are bright. The high contrast between the brain tissues and CSF allows for better visualization of the brain tissues and any pathology that may be present. FLAIR is a standard sequence in many neuroimaging protocols, providing high-quality images of brain tissues, and is useful in diagnoses of stroke, tumor, and multiple sclerosis.

A brief summary of FLAIR generation is as follows: first, an inversion pulse is applied to invert the magnetization of all the protons in the tissue. Then, a delay is introduced to allow the inverted magnetization to reach its steady state which is set to long enough for the

CSF to return to the null point [BAD96].

In stroke diagnoses, FLAIR MRI can be used to detect the presence of ischemic lesions, which appear as areas of increased signal intensity due to the accumulation of water in the affected brain tissue. FLAIR images can also be used to distinguish between acute and chronic strokes, as acute ischemic lesions typically appear hyperintense on FLAIR images within the first few hours of the stroke onset. In addition, FLAIR MRI can help to identify other conditions that may mimic stroke, such as brain tumors, abscesses, or inflammatory diseases.

2.2.1.3 Perfusion Weighted Imaging

Perfusion MRI is a non-invasive advanced imaging technique that is widely used in the assessment of various neurological disorders, including stroke. Perfusion MRI measures blood flow in the brain by tracking the passage of a contrast agent through the cerebral vasculature. Dynamic Susceptibility Contrast-enhanced (DSC) MR perfusion, Dynamic Contrast-Enhanced (DCE) MR perfusion, and Arterial Spin Labeling (ASL) are among the common MR perfusion techniques. Note that DCE is T1-weighted, DSC is T2*-weighted, and ASL requires no contrast agents.

For example, for a DSC perfusion MRI, the process of perfusion MR generation is as follows: first, a contrast agent, such as Gadolinium-based contrast, is injected into the patient's bloodstream. Then, a series of images are acquired over time as the contrast agent passes through the brain tissue. As the contrast agent passes through tissues, it induces a reduction of T2* signals in the nearby water molecule due to the susceptibility effect of the contrast agent that distorts the local magnetic field. By analyzing the changes in time-signal intensity in the images through relaxation, various perfusion parameters can be calculated, including cerebral blood flow (CBF), cerebral blood volume (CBV), mean transit time (MTT), and time to peak (TTP). These perfusion parameters provide valuable information about the hemodynamic status of the brain tissue and can be used to identify

regions of the brain that are ischemic [ESN13].

Perfusion MRI is particularly useful in the evaluation of stroke patients, as it can help to determine the extent of the ischemic damage and the potential for tissue recovery. In acute ischemic stroke, perfusion MRI can identify areas of the brain that are salvageable and may benefit from reperfusion therapy, such as thrombectomy or tissue plasminogen activator (tPA). Perfusion MRI can also be used to monitor the progression of stroke and to evaluate the effectiveness of treatment over time.

In addition to its clinical applications in stroke, perfusion MRI has also been used in research settings to investigate various aspects of cerebral physiology and pathophysiology. For example, perfusion MRI has been used to study the effects of aging on cerebral blood flow and to identify biomarkers of neurodegenerative diseases such as Alzheimer's disease.

Overall, perfusion MRI is a powerful imaging tool that provides valuable information about the hemodynamic status of the brain tissue and is widely used in the assessment and management of stroke patients.

2.2.2 Computed Tomography

Computed tomography (CT) is an imaging technique that uses X-rays and computer processing to produce detailed images of the body's internal structures, including the brain. CT is the first-line used imaging modality in the diagnosis and management of stroke patients.

In CT, an X-ray source rotates around the patient, and detectors measure the radiation that passes through the body. The resulting data is processed by a computer to generate cross-sectional images of the body. Both contrast-enhanced and non-contrast CT (NCCT) scans are widely used. NCCT can be rapidly acquired with a relatively low dose of radiation. NCCT is particularly useful in the initial image review of stroke patients because it can provide a rapid and accurate diagnosis of acute ischemic stroke and other stroke-related conditions, such as hemorrhagic stroke. Guidelines recommend at least 50% stroke patients

undergo NCCT within 25 minutes of arrival to determine whether there is ICH or a large hypo-attenuating infarct [PRA19].

CT angiography (CTA) is a type of CT that is used to visualize the blood vessels in the brain and to identify the location of the clot. CT angiography should be performed immediately after NCCT scan to identify intracranial LVOs in patients with acute MCA or intracranial internal carotid artery (ICA) syndromes for EVT preparation [PVG19]. When injected iodine-based contrast agent, a series of CT images can be acquired at different time intervals and reconstructed by maximum intensity projection (MIP), as well as generating CT perfusion (CTP) images that can be further processed similarly as MR Perfusion to acquire perfusion parameter maps.

2.3 Deep Learning

2.3.1 Convolutional Neural Network

Convolutional neural network (CNN) is a special type of Artificial Neural Network (ANN) that excels at many computer vision tasks in the recent decade. The "neocognitron" proposed by Fukushima [Fuk80] is considered the pioneer of CNN. Modern CNNs were established by introducing back-propagation by [RHW86]. Waibel *et al* proposed the first one-dimensional CNN called Time Delay Neural Network (TDNN) [WHH89]. LeCun *et al* developed the LeNet CNN architecture for handwritten digit recognition from U.S. Postal Service and defined the basic components of a CNN which contains convolutional layers, pooling layers, and fully connected layers [LBD89]. 2.3 shows an example of a CNN.

Convolutional layers is the backbone of a CNN. It performs a mathematical operation called convolution, which involves sliding parameterized kernels (also known as filters) that convolve over the input data to extract features and generate new feature maps. This process helps the model learn local patterns, such as edges, corners, or textures, in the case of images. Convolutional layers can have multiple filters, enabling the model to learn various

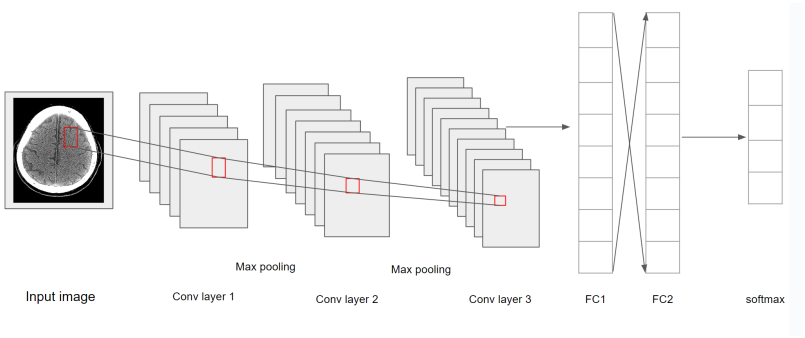


Figure 2.3: a simple 3-layer CNN architecture

features from the input data [KSH17]. Two advantages of CNN compared to ANN are 1) local connectivity means that each neuron is only connected to a small number of neurons or units from the previous layer. It helps reduce the number of parameters and the model convergence; 2) Weights of different filters within a feature map are shared across different spatial locations, further reducing the number of parameters.

Pooling layers reduce the spatial dimensions of the feature maps by aggregating representations within a small region into one, effectively down-sampling the features. Common pooling operations are mean or max pooling. The pooling layer reduces the number of parameters required by the following layers and controls overfitting by passing through summary statistics to the next layer rather than direct weights.

Fully connected layers connect all the neurons from one layer to the next layer. Fully connected layers have significantly more parameters than Convolutional layers. Fully connected layers are usually attached at the end of CNN before the final output layer.

In addition, activation layers [HKM22] (e.g. Sigmoid, Tanh, ReLu, LeakyReLu), batch normalization [IS15], and dropout [SHK14] are added to CNN as intermediate layers to prevent overfitting and accelerate model convergence. CNNs are built by stacking convolutional layers and other intermediate layers to extract imaging features from low level to high level in a hierarchical way. Shallower layers extract low-level patterns such as edge and texture features while in the later deep layers, the model extracts high-level features such as global

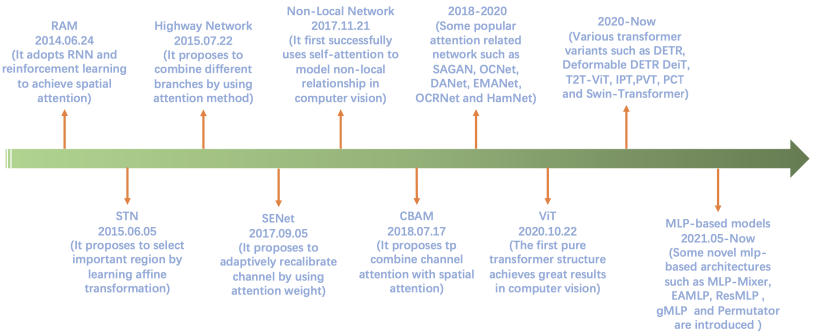


Figure 2.4: Timeline of attention mechanism development. This figure is credited to [GXL22]

information when the image is downsampled to be captured by the model in fewer parameters and a larger receptive field. CNNs have been improved over the years with notable architectures including AlexNet [KSH17], VGG [SZ14], ResNet [HZR16], DenseNet [HLV17], EfficientNet [TL19], and ConvNeXt [LMW22]. Although in recent years, transformer architecture shows more promising performance on large-scale datasets, the development by ConvNeXt shows that with modern modifications, CNN still has strong performance.

2.3.2 Attention Mechanism

The attention mechanism enables models to selectively focus on certain parts of the input data while processing it, essentially mimicking the human ability to pay attention to specific aspects of an input. In computer vision, an attention mechanism is a technique used to focus the processing of an image or video sequence on specific regions or objects of interest. The attention mechanism allows a model to selectively attend to relevant parts of the input while filtering out irrelevant information, which can improve the accuracy and efficiency of the model. 2.4 shows a brief summary of key development in attention in computer vision [GXL22].

The attention mechanism was first introduced by Bahdanau *et al* in 2014 [BCB14] to address the limitations of fixed-length context vectors in sequence-to-sequence models for machine translation. It allows the model to dynamically weigh the importance of input

elements and assign different levels of focus to them, resulting in improved performance. Attention mechanisms can be applied to different types of computer vision tasks, such as object recognition, segmentation, and detection. In object recognition, attention mechanisms can be used to focus on specific parts of an image that contain the object of interest, while ignoring other distracting elements in the scene. In segmentation, attention mechanisms can be used to highlight the boundaries between objects in an image, which can improve the accuracy of the segmentation process. In detection, attention mechanisms can be used to highlight the regions of an image that contain potential objects, which can improve the speed and efficiency of the detection process.

Attention mechanisms can be categorized into two major types: Hard attention and soft attention. Soft attention is most commonly used in literature, which calculates a probability distribution range from 0 to 1 that provides a continuous weighting across inputs [XBK15]. Soft attention assigns weights to different parts of the input, thus all the weights are smoothed and differentiable and can be learned through back-propagation. These weights are then used to compute a weighted sum of the input features, which emphasizes the relevant parts of the input while suppressing the irrelevant parts. In the hard attention mechanism, the weights follow the Bernoulli distribution which takes values of 0 or 1. Therefore, the weights are not differentiable and the gradients cannot be updated by back-propagation [MHG14].

Attention mechanisms can also be categorized by data domain. Channel Attention generates attention masks across the channel dimension and uses them to select important channels. Squeeze-and-Excitation network (SENet) is among the representative works of this type [HSS18]. SENet designed a squeeze-and-excitation (SE) block that is composed of a squeeze module and an excitation module. Using global average pooling, the squeeze module collects the spatial information and the excitation module collects channel-wise relationship information. The attention vector is generated by using FC layers and non-linear activation layers. Spatial Attention guides the model to focus on important regions while suppressing unrelated regions. RAM [MHG14] uses Recurrent Neural Network (RNN) [HS97]

and Reinforcement Learning (RL) [SMS99] to model the attention mechanism as a sequential decision process to let the model pay attention to the region of interest. Self-attention is the foundation of the successful Non-local Network [WGG17] and later becomes the foundation of transformer [VSP17] architecture which becomes the backbone for state-of-the-art Natural language processing (NLP) and computer vision (CV) models. Due to the locality characteristics of CNN, the global understanding capability is limited by narrow receptive fields. To overcome this, self-attention is introduced to CV. The self-attention can be defined as:

1. Linear projections of input embeddings:

$$QueryQ = W^Q * X \tag{2.2}$$

$$KeyK = W^K * X \tag{2.3}$$

$$ValueV = W^V * X \tag{2.4}$$

Where X represents the input sequence embeddings, W^Q, W^K, and W^V are learnable weight matrices, and Q, K, and V are the projected query, key, and value matrices, respectively.

2. Scaled dot-product attention:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.5}$$

Where QK^T represents the dot product between the query and key matrices, d_k is the dimension of the key vectors, and the result is normalized using the square root of this dimension. The softmax function is applied to the result, followed by the matrix multiplication with the value matrix V.

However, the main drawback of self-attention in direct application in CV is its complexity due to its quadratic scaling by unit-to-unit computation.

2.3.3 Vision Transformers

Vision Transformer, also known as ViT, is a state-of-the-art Deep Learning (DL) architecture for both NLP and CV tasks. It was proposed by researchers at Google in 2020 [DBK20] and is based on the Transformer architecture that was originally developed for natural language processing (NLP) tasks [VSP17].

The key innovation of ViT is the use of self-attention mechanisms to capture the long-distance relationships between different parts of the image. In traditional convolutional neural network (CNN) architectures, the input image is processed by a series of convolutional layers that extract local features, which are then aggregated by pooling layers to create a global representation of the image. In contrast, ViT uses a sequence of attention blocks, each of which attends to different parts of the input image, allowing it to capture both local and global relationships.

The ViT architecture consists of three main components: a patch embedding layer, a transformer encoder, and a classification head. In the patch embedding layer, the input image is divided into a grid of fixed-size patches, which are then projected into a high-dimensional feature space. The transformer encoder consists of a series of attention blocks, each of which applies self-attention to the input features, followed by a feedforward network that applies non-linear transformations. The classification head is a simple fully connected

layer that maps the final output of the transformer encoder to the class labels.

1. Input Image and Patch Embeddings:

Given an input image $I \in \mathbb{R}^{H \times W \times C}$ with height H, width W, and C channels, the image is divided into N fixed-size non-overlapping patches $P_i \in \mathbb{R}^{P_H \times P_W \times C}$, where P_H and P_W are the patch height and width, respectively. Each patch is then linearly embedded into a flat vector $x_i \in \mathbb{R}^D$, where D is the desired embedding dimension:

$$x_i = Flatten(P_i)W_E \tag{2.6}$$

Where $W_E \in \mathbb{R}^{(P_H \cdot P_W \cdot C) \times D}$ is a learnable embedding matrix.

2. Position Embeddings:

Position embeddings are added to the patch embeddings to incorporate the spatial information of each patch:

$$X = x_1 + p_1, x_2 + p_2, ..., x_N + p_N \tag{2.7}$$

Where $p_i \in \mathbb{R}^D$ are the learnable position embeddings.

3. Adding a Learnable Classification Token:

A learnable classification token $c \in \mathbb{R}^D$ is prepended to the sequence of patch embeddings:

$$X' = [c; X] \tag{2.8}$$

4. Transformer Layers:

The modified input sequence X' is then fed through L layers of the Transformer architecture. Each layer consists of multi-head self-attention and position-wise feed-forward networks, as well as layer normalization:

$$Z^{(l)} = LayerNorm(X^{(l)} + MultiHeadAttention(X^{(l)})) \tag{2.9}$$

$$X^{(l+1)} = LayerNorm(Z^{(l)} + FFN(Z^{(l)})) \tag{2.10}$$

Where l represents the layer index, and $X^{(l)}$ and $Z^{(l)}$ are the input and output of the l-th layer, respectively.

5. Classification:

After the final Transformer layer, the classification token is extracted, and a linear layer followed by a softmax function is applied to produce the probability distribution over the target classes:

$$y = Softmax(c^{(L+1)}W_C + b_C) \tag{2.11}$$

Where $c^{(L+1)}$ is the classification token after the final Transformer layer and W_C and b_C are learnable weight and bias parameters for the classification layer.

ViT has achieved state-of-the-art performance on several benchmark datasets, including ImageNet, COCO, and CIFAR, with significantly fewer parameters than previous CNN-based models. It shows that by using large-scale training, the model overcomes the CNNs inductive bias characteristics that help the model generalize. Its success has demonstrated the potential of self-attention mechanisms for image classification tasks and has opened up new avenues for research in DL for CV.

2.3.3.1 Swin Transformers

Swin Transformer (Swin-T) is a recently proposed DL architecture for image recognition tasks, which was introduced by researchers from Microsoft Research Asia in 2021 [LLC21]. Swin-T is an extension of the ViT architecture, which improves upon its limitations and has achieved state-of-the-art results on several benchmark datasets.

The main innovation of the Swin-T is its hierarchical architecture, which allows for the

efficient processing of images at multiple scales. In ViT, the input image is divided into fixed-size patches, which are processed independently by the transformer network. However, this approach can be limiting for high-resolution images, as it requires a large number of patches and therefore a large number of parameters.

In contrast, Swin-T uses a hierarchical approach, where the input image is first divided into a small number of patches, which are processed by a local transformer network. The output of this network is then grouped into larger patches, which are processed by a higher-level transformer network. This process is repeated several times, creating a hierarchy of transformer networks that can efficiently process images at multiple scales.

Another key innovation of Swin-T is its use of shift operations, which enable it to capture spatial relationships between neighboring patches. In the original ViT, the self-attention mechanism is used to capture relationships between different parts of the input but does not explicitly model the spatial structure of the input. The shift operation in Swin Transformer allows the model to capture spatial relationships between patches without the need for additional convolutional layers.

The detailed Swin Transformer architecture is as follows: 1. Image Patch Partitioning and Embeddings:
Similar to ViT, the input image $I \in \mathbb{R}^{H \times W \times C}$ is divided into non-overlapping patches $P_i \in \mathbb{R}^{P_H \times P_W \times C}$. Each patch is linearly embedded into a flat vector $x_i \in \mathbb{R}^D$:

$$x_i = Flatten(P_i)W_E \tag{2.12}$$

Where $W_E \in \mathbb{R}^{(P_H \cdot P_W \cdot C) \times D}$ is a learnable embedding matrix.

2. Local Window Partitioning:
The patches are then partitioned into non-overlapping local windows of size $M \times M$. For an image with $H \times W$ patches, there are $\frac{H}{M} \times \frac{W}{M}$ local windows.

3. Shifted Window-based Self-Attention:

Shifted window-based self-attention is applied to each local window. In each layer, the local windows are shifted by a half-window size in both horizontal and vertical directions. The self-attention mechanism can be described as follows:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.13}$$

Where QK^T is the dot product between the query and key matrices, and d_k is the dimension of the key vectors.

4. Hierarchical Processing:

The Swin Transformer processes images in a hierarchical manner by successively merging adjacent non-overlapping patches. This is achieved by applying a patch merging layer, which consists of a linear transformation followed by element-wise addition of the position embeddings:

$$X' = LayerNorm(X + PositionEmbedding) \tag{2.14}$$

$$Z = X'W_M \tag{2.15}$$

Where $W_M \in \mathbb{R}^{D \times D'}$ is a learnable weight matrix, and D' is the desired output dimension.

5. Transformer Layers:

Similar to the Vision Transformer, the Swin Transformer has several layers of multi-head

self-attention and position-wise feed-forward networks, as well as layer normalization:

$$Z^{(l)} = LayerNorm(X^{(l)} + MultiHeadAttention(X^{(l)})) \tag{2.16}$$

$$X^{(l+1)} = LayerNorm(Z^{(l)} + FFN(Z^{(l)})) \tag{2.17}$$

Where l represents the layer index, and $X^{(l)}$ and $Z^{(l)}$ are the input and output of the l-th layer, respectively.

6. Classification:

After the final Transformer layer, global average pooling is applied to the output feature map, followed by a linear layer and a softmax function to produce the probability distribution over the target classes:

$$y = Softmax(GlobalAveragePooling(X^{(L+1)})W_C + b_C) \tag{2.18}$$

Where W_C and b_C are learnable weight and bias parameters for the classification layer.

Swin Transformer exceeds the performance of ViT on various tasks without the requirement of large-scale training data by introducing back inductive bias and translational invariance that is ignored by ViT.

2.3.4 Self Supervised Learning

Self-Supervised Learning (SSL) is a subfield of Unsupervised Learning. SSL involves training a model on unlabeled data by solving auxiliary tasks, often referred to as pretext tasks, where

the model is provided with a set of inputs and is tasked with predicting some aspect of the input data, without any explicit labels or annotations. The learned representations can then be used for downstream tasks, such as classification or regression, with little or no fine-tuning.

One of the main challenges in Supervised Learning is the reliance on large amounts of labeled data, which can be expensive and time-consuming to obtain. SSL aims to address this issue by leveraging the vast amount of raw, unlabeled data available, allowing models to learn useful features and representations without the need for manual annotation. In SSL, the model is trained on a large amount of unlabeled data, such as images, text, or audio, which is much easier and cheaper to obtain than labeled data. By learning to predict some aspect of the input data, such as predicting the rotation, colorization, or context of an image, the model can learn meaningful representations of the input data, which can be used for a wide range of downstream tasks, such as image classification, object detection, and natural language processing.

SSL has shown promising results in a variety of domains and has been particularly successful in computer vision and natural language processing tasks. In NLP, it revolutionized the field along with Transformers. Masked language models such as BERT [DCL18] and GPT series [BMR20] are the two main representative SSL frameworks. BERT predicts the missing word in a sentence while GPT predicts the next word after seeing a sequence. In CV, SSL can be categorized into two groups: Contrastive Learning and Generative Learning and we will cover these two topics in the following sections.

2.3.4.1 Contrastive Learning

Contrastive Learning is a type of SSL that involves training a model to distinguish between pairs of similar and dissimilar inputs. In Contrastive Learning, the model is trained on a set of inputs, such as images or text, which are divided into pairs of similar and dissimilar samples.

The objective of Contrastive Learning is to learn a representation space where similar samples are mapped close together, while dissimilar samples are mapped far apart. This is achieved by minimizing the distance between similar samples and maximizing the distance between dissimilar samples, using a loss function such as contrastive loss or triplet loss. By learning to distinguish between similar and dissimilar inputs, the model can learn to capture the underlying structure and regularities in the input data, which can be useful for a wide range of downstream tasks, such as image classification, object detection, and natural language processing.

SimCLR [CKN20] and MoCo [HFW20] are the two most popular approaches for Contrastive Learning. In MoCo, the design leverage instance discrimination by substantially increasing negative samples through a Momentum-updated encoder that stores a dynamic dictionary. The MoCo framework consists of

Contrastive Learning shows that discriminative models can learn useful feature representations, breaking the longtime belief that the generative model is the only choice for representation learning. The MoCo framework consists of two encoders, f_q and f_k, which are used to compute the query and key representations, respectively. f_q is the main encoder, which is updated by backpropagation, while f_k is the momentum encoder, which is updated as a moving average of f_q. The momentum update is defined as follows:

$$f_k \leftarrow m \cdot f_k + (1 - m) \cdot f_q \tag{2.19}$$

where $m \in (0, 1)$ is the momentum coefficient. The higher the value of m, the smoother the update process.

In MoCo, the contrastive loss function is the InfoNCE loss, which aims to maximize the mutual information between the query and key representations. Given a query $q = f_q(x_q)$ and a set of keys $K = k_1, \ldots, k_N$, with $k_+ = f_k(x_+)$ being the positive key and the remaining

keys being negatives, the InfoNCE loss is defined as:

$$L(q, K) = -\log \frac{\exp(q^\top k_+/\tau)}{\sum_{i=1}^{N} \exp(q^\top k_i/\tau)} \tag{2.20}$$

where τ is the temperature hyperparameter to control the sharpness of the distribution. A higher value of τ produces a softer distribution, making the model less sensitive to small differences between similarities, while a lower value results in a more focused distribution.

However, MoCo v1's positive sample pair does not use transformation or augmentation, making the model too easy to distinguish. In SimCLR, the authors demonstrated the importance of hard positive samples by introducing augmented views of the same data samples to construct positive and negative pairs in 10 forms such as crop and resize, color distort, flipping, noise, and rotation. The details are as follows: Given an input image x, we generate two augmented views x_i and x_j using the data augmentation module. These views are then passed through the base encoder $f(\cdot)$ and the projection head $g(\cdot)$ to compute the representations $z_i = g(f(x_i))$ and $z_j = g(f(x_j))$. SimCLR uses a contrastive loss called NT-Xent (Normalized Temperature-Scaled Cross-Entropy) loss. The positive pair in this case is the two augmented views of the same image, while the negative pairs are other augmented views in the same batch. The NT-Xent loss for a positive pair (i, j) is given by:

$$\mathcal{L}i, j = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum k = 1^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2.21}$$

Here, $\text{sim}(z_i, z_j) = \frac{z_i^\top z_j}{|z_i||z_j|}$ is the cosine similarity between the representations z_i and z_j, τ is a temperature hyperparameter. N is the minibatch size. However, SimCLR returns to the old large-scale negative samples approach that is hard to train unless the batch size is very large (e.g. 8196). MoCo v2 [CFG20] adopt the findings in SimCLR and further improve the performance. Many more following works aim to further improve the performance of Contrastive Learning. BYOL [GSA20] is one of the latest popular choices in Contrastive SSL. There are still several major limitations. For instance, the COCO object detection task

does not benefit from contrastive SSL. Although a lot of efforts have been made to reduce batch size, it is still a lot bigger than the batch size usually seen in the medical domain.

2.3.4.2 Masked Image Model

Masked Autoencoder (MAE) [HCX21], is the first Masked Image Model (MIM) framework proposed to use along with ViT for SSL. MAE is based on Autoencoder (AE), which is a type of generative modeling, as opposed to Contrastive Learning, which uses discriminative modeling.

Like other AE architectures, the MAE consists of an encoder network and a decoder network. The encoder network maps the input data into a lower-dimensional latent space, while the decoder network maps the latent representation back into the original input space. The objective of the AE is to minimize the reconstruction error between the initial input and the reconstructed output.

The key difference between an MAE and a standard AE is that the input data to the MAE is partially masked or corrupted by replacing some of the input values with a mask value. Instead of attempting to reconstruct the entire input, MAE focuses on reconstructing a masked or corrupted portion of the input, thus encouraging the model to capture the underlying structure and semantics of the data. The mask value can be chosen randomly or systematically, such as by setting a random subset of pixels to zero in an image.

Let's denote the input data as $x \in \mathbb{R}^d$, and let $M(x)$ be a masking function that produces a corrupted version of the input, with some portion masked or removed. The encoder, denoted as $f_{\theta_e}(\cdot)$, maps the corrupted input to a latent representation $z \in \mathbb{R}^l$:

$$z = f_{\theta_e}(M(x)) \tag{2.22}$$

The decoder, denoted as $f_{\theta_d}(\cdot)$, reconstructs the masked portion of the input from the latent representation:

$$\hat{x} = f_{\theta_d}(z) \tag{2.23}$$

The goal of the Masked Autoencoder is to minimize the reconstruction error between the original input and the reconstructed input. This can be measured using a loss function such as the mean squared error (MSE). The loss function can be written as:

$$\mathcal{L}(\theta_e, \theta_d) = \sum_{i \in \mathcal{M}} \mathcal{D}(x_i, \hat{x}_i) \tag{2.24}$$

where $\mathcal{M}$ is the set of indices corresponding to the masked elements in the input, and $\mathcal{D}$ is a distance metric, such as the squared error $(x_i - \hat{x}_i)^2$ or the cross-entropy $-x_i \log \hat{x}_i - (1 - x_i) \log(1 - \hat{x}_i)$, depending on the nature of the input data.

The MAE can be used for a variety of tasks, such as image denoising, inpainting, and anomaly detection. It has also been used as a pre-training step for downstream tasks, such as classification and clustering, to learn more effective representations of the input data. Recently, there are also other popular MIM frameworks. BEiT [BDP21] is a slightly earlier work than MAE that first uses a discrete variational autoencoder (dVAE) to tokenize image patch to discrete visual tokens, then use masked strategy to do reconstruction. MAE demonstrated that the image patch tokenize is not necessary by direct reconstructing the RGB values of the image patches. At about the same time as MAE work, SimMIM [XZC21] was proposed with a reduced decoder that contained only one linear layer. Unlike MAE which only input masked region, SimMIM input all the image patches. More importantly, SimMIM is designed to work with SwinT, which is a hierarchical structure that a simple patching strategy like MAE cannot process.

2.4 Super Resolution

Super-resolution (SR) refers to the task of generating a high-resolution (HR) image or video from one or more low-resolution (LR) input images or frames. The goal is to recover missing

details and enhance the quality of the LR images, making them visually similar to HR images.

SR is an important problem in image and video processing, with applications in a wide range of domains such as medical imaging, satellite imaging, surveillance, and consumer electronics.

There are different approaches to SR, including interpolation-based methods, which use interpolation techniques such as the nearest neighbor, bilinear pooling, or bicubic to increase the resolution of an image, and learning-based methods, which learn a mapping from LR images to HR images using DL.

Learning-based SR methods have shown promising results in recent years, particularly with the development of CNN architectures. The primary goal is to learn a mapping function between the low-resolution input image and its high-resolution counterpart, typically through a dataset containing pairs of low- and high-resolution images using different loss functions, such as mean squared error or perceptual loss. Using SRCNN [DLH15] as an example, the problem can be formulated as: Given a low-resolution input image $Y \in \mathbb{R}^{W \times H \times C}$, where W, H, and C are the width, height, and number of channels, respectively, the objective of the SRCNN is to learn a mapping function F that reconstructs the high-resolution image $X \in \mathbb{R}^{W' \times H' \times C}$:

$$X = F(Y; \Theta) \tag{2.25}$$

Here, Θ denotes the learnable parameters of the SRCNN.

The training process involves minimizing a loss function, which measures the difference between the ground-truth high-resolution image X and the reconstructed image $F(Y; \Theta)$. A common loss function used for this task is the mean squared error (MSE) loss:

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} |X^{(i)} - F(Y^{(i)}; \Theta)|^2 \tag{2.26}$$

Here, n denotes the number of training samples, and $X^{(i)}$ and $Y^{(i)}$ are the ground-truth high-resolution image and its corresponding low-resolution input for the i-th sample, respectively.

There are other more advanced DL models for super-resolution, such as the Enhanced Deep Residual Networks for Single Image Super-Resolution (EDSR) [LSK17], the Very Deep Super-Resolution (VDSR) [KLL16], and the Generative Adversarial Networks-based super-resolution (SRGAN) [LTH17]. These models employ more complex architectures and loss functions to improve the quality of the reconstructed high-resolution images. SR can also be applied to video frames, with methods such as temporal super-resolution, which uses information from multiple frames to generate HR frames, and video SR, which generates HR videos from LR videos.

CHAPTER 3

Using 2D and 3D Attention CNN and Self-supervised Learning to Determine Acute Ischemic Stroke Onset Time with Pretreatment MRI

3.1 Introduction

Self-supervised learning attempts to use unlabeled data to pretrain a model and fine-tunes it for downstream tasks. However, pretraining on the same labeled dataset with some pretext objections may also help the model converge on the downstream task without introducing large-scale external unlabeled datasets. In this chapter, we will first introduce our early exploration of self-supervised learning, namely "intra-domain task adaptive transfer learning" to predict stroke onset time using pretreatment MRI. The work described in this chapter is in press as Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time [ZPN21b] and Identifying acute ischemic stroke patients within the thrombolytic treatment window using deep learning [PZN22]. A follow-up study on DWI-FLAIR mismatch is also covered in appendix [PZN21].

3.2 Overview

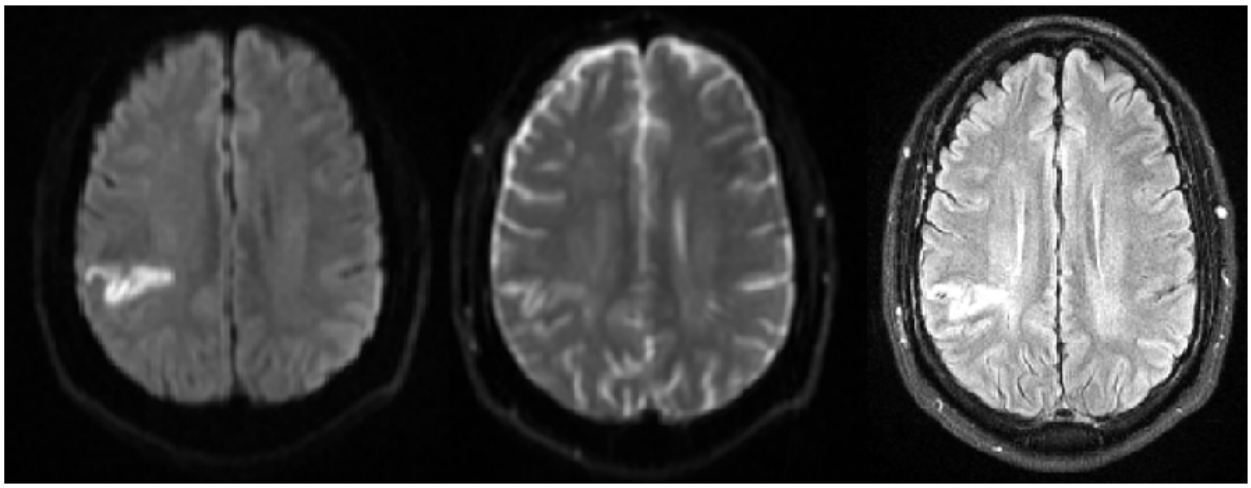
Acute Ischemic Stroke (AIS) is a type of cerebrovascular disorder responsible for approximately 2.7 million deaths globally each year [Ben19]. he approach to AIS treatment is significantly influenced by the time since stroke onset (TSS); current clinical guidelines ad-

vocate for thrombolytic therapies for AIS patients who present within a 4.5-hour window from stroke onset and endorse endovascular thrombectomy for those presenting within a 24-hour window. A sizable proportion of AIS cases, up to 25%, occur without a clearly defined TSS [TFO14, UFZ18]. This could be due to strokes that are unwitnessed, strokes that occur during sleep, or situations where patient reporting is unreliable. For this particular group of patients, the most recent American Heart Association (AHA) guidelines suggest the use of specific MRI sequences to determine patient eligibility for thrombolytic therapy [PRA19].

Following the WAKE-UP trial [TSB18], which used DWI-FLAIR mismatch to select patients for extending the time window for intravenous thrombolysis, the use of MRI (DWI-FLAIR mismatch) is now recommended (level IIa) to identify unwitnessed AIS patients who may benefit from thrombolytic treatment [PRA19]. Specifically, diffusion-weighted imaging (DWI) displays the increased signal in ischemic areas within minutes of stroke occurrence, while fluid-attenuated inversion recovery (FLAIR) imaging can show fluid accumulation after a few hours [EBS18], as shown in Figure 3.1. A DWI-positive, FLAIR-negative mismatch can identify stroke lesions that could benefit from the administration of thrombolytics. However, assessing this mismatch is subject to high variability compared across multiple readings and/or radiologists [Tho11]. Hence, the ability to accurately ascertain stroke onset solely through imaging could expand the pool of patients eligible for thrombolytic treatments, potentially leading to enhanced patient outcomes.

Several Machine Learning (ML) approaches have been used for automated determination of stroke onset time. These approaches typically involve creating features — either hand-crafted, radiomics-based, or derived through Deep Learning (DL) — from clinical reports or imaging data. These features are then utilized as input for a diverse range of ML models. [HSZ19, HSE17, LLH20]. The extraction of these features has traditionally been dependent on predetermined regions of interest, which are usually identified through image thresholding or using a parameter map. By focusing solely on these immediate regions, we may overlook crucial imaging traits within the surrounding area. Considering the interconnected nature

DWI-FLAIR Match Case



DWI-FLAIR Mismatch Case

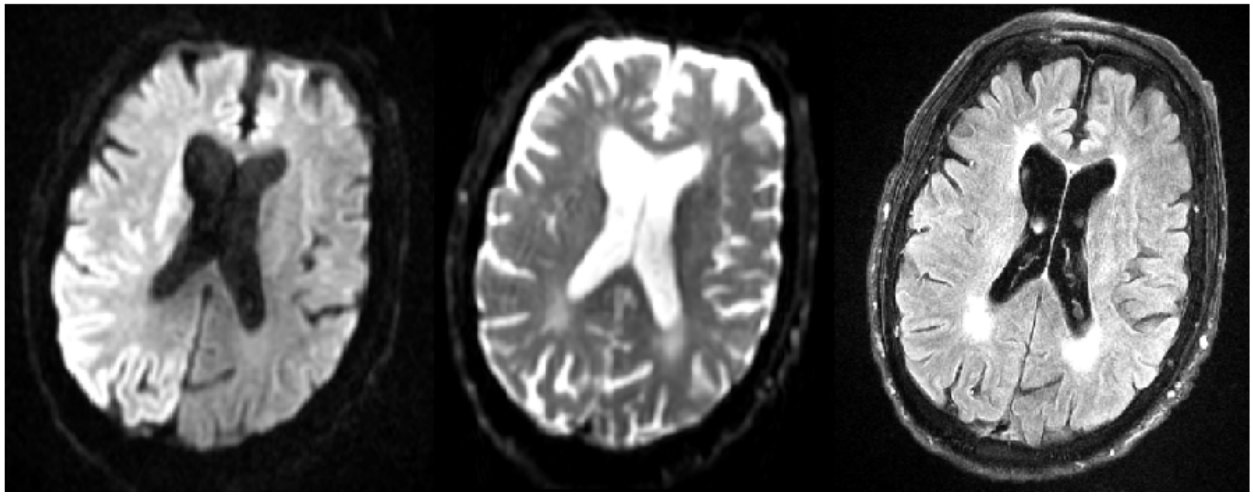


Figure 3.1: Sample cases of DWI-FLAIR Mismatch. Sequences from left to right: DWI b1000, DWI B0, FLAIR

of cerebral blood flow, these characteristics could hold vital information in determining TSS [Ban11]. Moreover, prior approaches have often applied stringent exclusion criteria, based on either the location of the stroke or image-related factors associated with preprocessing. As a result of these strict criteria, a substantial number of patients, up to 40% in some studies, were deemed ineligible for evaluation [LLH20].

DL models have excelled in medical imaging for segmentation and classification tasks [Shi16, MNA16, CSH20, TZY22, WHM18]. Specifically, convolutional neural networks (CNNs) have produced state-of-the-art results even in small datasets that are commonly seen in medical imaging research [Lit17]. Convolutional operations, which collectively analyze neighborhood pixels across multiple layers, can be conducted in either two or three dimensions. Although a broad spectrum of 2D Convolutional Neural Networks (CNNs) have been applied to medical imaging tasks, 3D CNNs provide the additional benefit of incorporating information along the vertical dimension (z-axis). However, these potential benefits of 3D convolutions come at the expense of increased model complexity, which typically necessitates larger volumes of data and greater computational power for effective training.

Due to the large number of parameters in a deep neural network, a high volume of data is typically required for training. For particularly complex classification tasks, transfer learning has proven effective in not only reducing computational demands and shortening the time required for model convergence but also enhancing performance, when compared to training models entirely from scratch [PY10]. Transfer learning generally follows a two-step process: initially, a model is trained on one dataset, then it's further refined on another dataset for a different task. Cross-domain transfer learning involves training on data from a source domain, and using those learned weights in a model trained on data from a different target domain [WKW16], e.g., from the natural image domain to the medical image domain or from the CT modality domain to the MR modality domain. Many DL approaches applied to medical images have used established architectures pre-trained on large natural image datasets such as ImageNet [RDS15] and refined the model to domain-specific tasks. This

method is believed to enhance model convergence and leverage the low-level features, initially learned on a large-volume dataset, for use on a smaller dataset. This strategy is especially pertinent to medical image models, given the significant costs associated with acquiring sufficient volumes of medical data. However, the differences in natural images and those in the medical domain limit the wide applicability of this method, likely due to the over-parameterization of the original models [CTB19]. Efforts have been made to pretrain models on public medical datasets, but access to such medical datasets is still limited. Moreover, higher-level features of medical images vary significantly for different medical domains. To combat the limitations of cross-domain transfer learning and increase feature reuse across models, intra-domain transfer learning has been implemented for both natural image and medical image tasks [RZK19]. Commonly, a model is initialized in a self-supervised or unsupervised fashion. The advantage of this approach is that it does not require outside datasets or labels. However, even intra-domain pretraining may result in limited feature reuse beyond the first convolutional layer [VVM20]. A task-adaptive approach, which uses the same data set for pretraining and then refines the model using two different label sets, has been demonstrated to increase feature reuse and enhance performance [Elm93, BLC09]. However, this has not yet been applied in the medical image domain.

We propose an intra-domain task-adaptive transfer learning approach and implement it for TSS classification. The approach uses a multi-stage training schema, leveraging features learned by training on an easier task (stroke detection) to refine the model for a more difficult task (TSS classification). We developed both 2D and 3D CNN models to classify TSS, and we demonstrated our proposed transfer learning approach enhanced classification performance for both architectures when compared to other pretraining schemas, with our 2D model achieving the best performance for classifying TSS < 4.5 hours. We also showed that adding soft attention mechanisms during the latter stages further improved the performance. To offer clinical insight, we compared our model performance to both previously published methods and radiologist assessment of DWI-FLAIR mismatch. Our DL models were able

to achieve greater classification sensitivity while maintaining specificity achieved by expert neuroradiologists. By visualizing network gradients via Grad-CAM [SCD19], we illustrated that our pre-trained models were able to localize the stroke infarct more precisely than the models trained from scratch. To our knowledge, this is the first end-to-end, DL approach to classify TSS on a patient dataset with minimal exclusion criteria; moreover, our model exceeds the performance of previously reported state-of-art ML models.

3.3 Dataset and Preprocessing

A total of 422 patients treated for AIS at the UCLA Ronald Reagan Medical Center from 2011-2019 were included in this study. This work was performed under the approval of the UCLA Institutional Review Board (#18-000329). A patient was included if they were diagnosed with AIS, had a known stroke onset time, and underwent MRI prior to any treatment if given. Clinical parameters were gathered from imaging reports and the patient record, with demographic data summarized in Table 3.1. The study cohort had a median age of 70 (55-80) years, a mean National Institutes of Health Stroke Scale (NIHSS) score of 8(4-15), and 56% female. The median onset to MRI was 222(105-715.25) minutes. For performance evaluation, we used 64% for training (272), 16% for validation (68), and 20% (82) as a hold-out test set. In order to prevent information leakage across tasks, the same test set was used across a set of experiments. The training and testing sets had similar distributions of these clinical factors and TSS. For each patient in the test cohort, DWI-FLAIR mismatch was assessed independently by three senior neuroradiologists with full access to all sequences used in our model.

For each patient, the T2w(DWI b0), DWI(DWI b1000), and FLAIR imaging sequences were retrieved from the institutional picture archiving and communication system (PACS). All patients underwent MRI using a 1.5T or 3T echo-planar Siemens MR imaging scanner, performed with 12-channel head coils. The FLAIR images were acquired using a TR

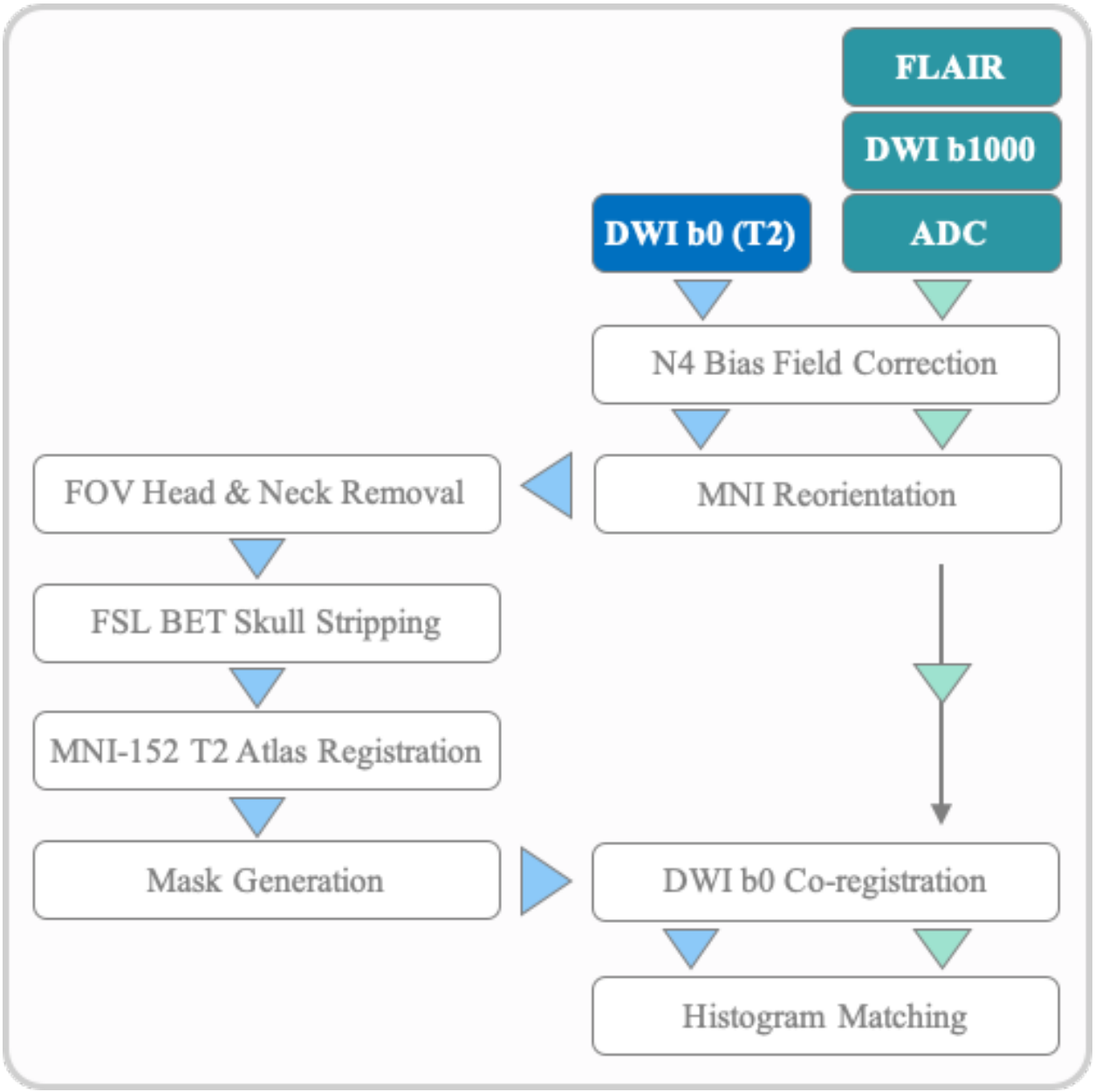


Figure 3.2: Preprocessing pipeline for patient series.

	Training Set (n = 340)	Test Set (n = 82)
Age (years)	70 (55-80)	68 (57-79)
Female	176 (52%)	46 (56%)
NIHSS	8 (4 - 16)	6.5 (2 - 18)
Onset to MRI (min)	210 (105-683)	230 (107-661)

Table 3.1: Patient cohort demographics. Numbers are n (%) or median (interquartile ranges). MRI indicates magnetic resonance imaging; NIHSS, National Institutes of Health Stroke Scale.

range of 8,000-9,000ms and a TE range of 88-134ms. The pixel dimension varied from 0.688x0.688x6.000mm to 0.938x0.938x6.500mm. The DWI images were acquired using a TR range of 4,000-9,000ms and a TE range of 78-122ms. The corresponding pixel dimensions varied from 0.859x0.859x6.000mm to 1.850x1.850x6.500mm. The DWI b0 sequence was used as a T2w proxy, as it denotes the first step of DWI acquisition with no diffusion attenuation, and the DWI here represents the sequence with a b-value equal to 1000. The rationale for using these sequences was: (1) T2w represents the anatomical image, so we theorized it might provide contrast information when input along with DWI and FLAIR sequences; (2) since our goal is to classify TSS, and the DWI-FLAIR mismatch is only a surrogate for this goal, extra anatomical imaging information could provide more features related to TSS; and (3) we used three sequences to mimic the RGB channels used in many image classification models, enabling us to compare our training schema to other pretraining approaches. After image retrieval, the sequences were fed into our automated preprocessing pipeline. First, N4 bias field correction [TAC10] was applied to all sequences. Then, each image series was reoriented to the T2w MNI-152 atlas [FEM09]. Next, the neck and skull were removed using FSL BET [SJW04]. The T2w sequence was registered using FSL FLIRT to a version of the T2w MNI-152 atlas that was resized to 224x224x26 using linear interpolation in order

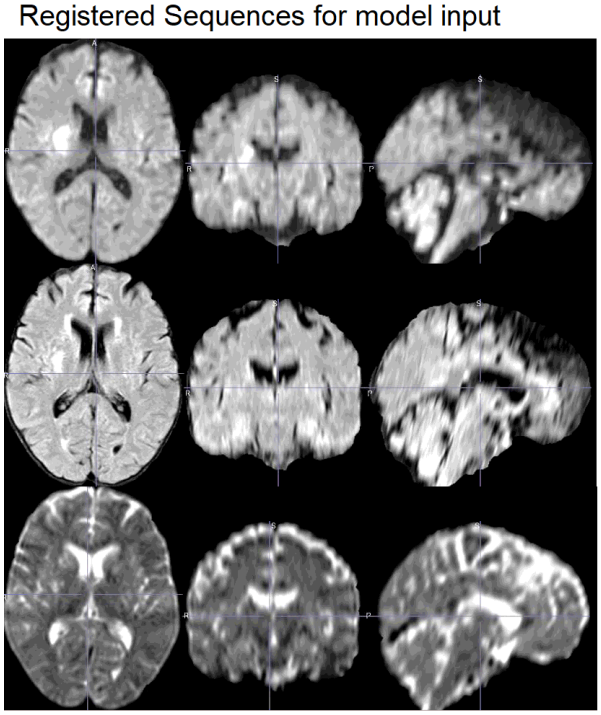


Figure 3.3: Sample case of Registered Output. Sequences from top to bottom: DWI(b1000), FLAIR, T2w(DWI b0).

to match the z dimension of the stroke sequences. After a second run of FSL BET was performed to remove remnant artifacts, the remaining sequences were co-registered to the T2w volume. Finally, the intensity was normalized, and histogram matching was performed using a reference study. A visual quality check was manually performed for all cases before the experiments. This data preprocessing pipeline is summarized in Figure 3.2 and sample output is shown in Figure 3.3.

3.4 Method

3.4.1 2D and 3D Model Architectures

We tested our intra-domain transfer learning schema on custom 2D and 3D architectures. The 2D network takes individual slices as input and feeds them through a convolutional backbone (ResNet-18) adapted from [HZR16] for feature extraction. To account for the large pixel input of an individual MRI slice, we also incorporated a soft attention gate into the architecture [SOS19]. This module uses the final and penultimate convolutional outputs to generate individual pixel weights which identify the most salient regions for the task. This attention module was refined during the TSS tasks later in training to avoid the possibility of convergence at a local minimum and precluding further optimization during model refinement [OSF18]. The attention module output and convolutional output were concatenated into a feature vector, which was then fed into a fully-connected layer to generate a single, slice-level output. To aggregate these slice-level predictions into an image-level prediction, we implemented a trainable weighting factor, ranging from 0 to 1, to assign a weight to each slice, and the slice-level outputs were summed in a weighted fashion, resulting in one probability label. The attention module and trainable weight factor ascribe pixel-level and slice-level importance that can be trained and optimized, which enables the model to localize to salient regions.

Given the 3D anatomical information in our dataset, we also evaluated a 3D model architecture. Our 3D model used the encoder part of the 3D U-Net as the model backbone [CAL16]. U-Net [RFB15], like ResNet, uses connections between layers for model training and also has been widely used in medical image research. Our 3D approach also used soft attention modules at the 128- and 256-channel intermediate outputs in the encoder part of 3D U-Net in order to allow the network to capture relevant information in the early stages of classification. Training a 3D CNN model from scratch does not necessarily yield better performance than 2D models due to the higher number of parameters and the

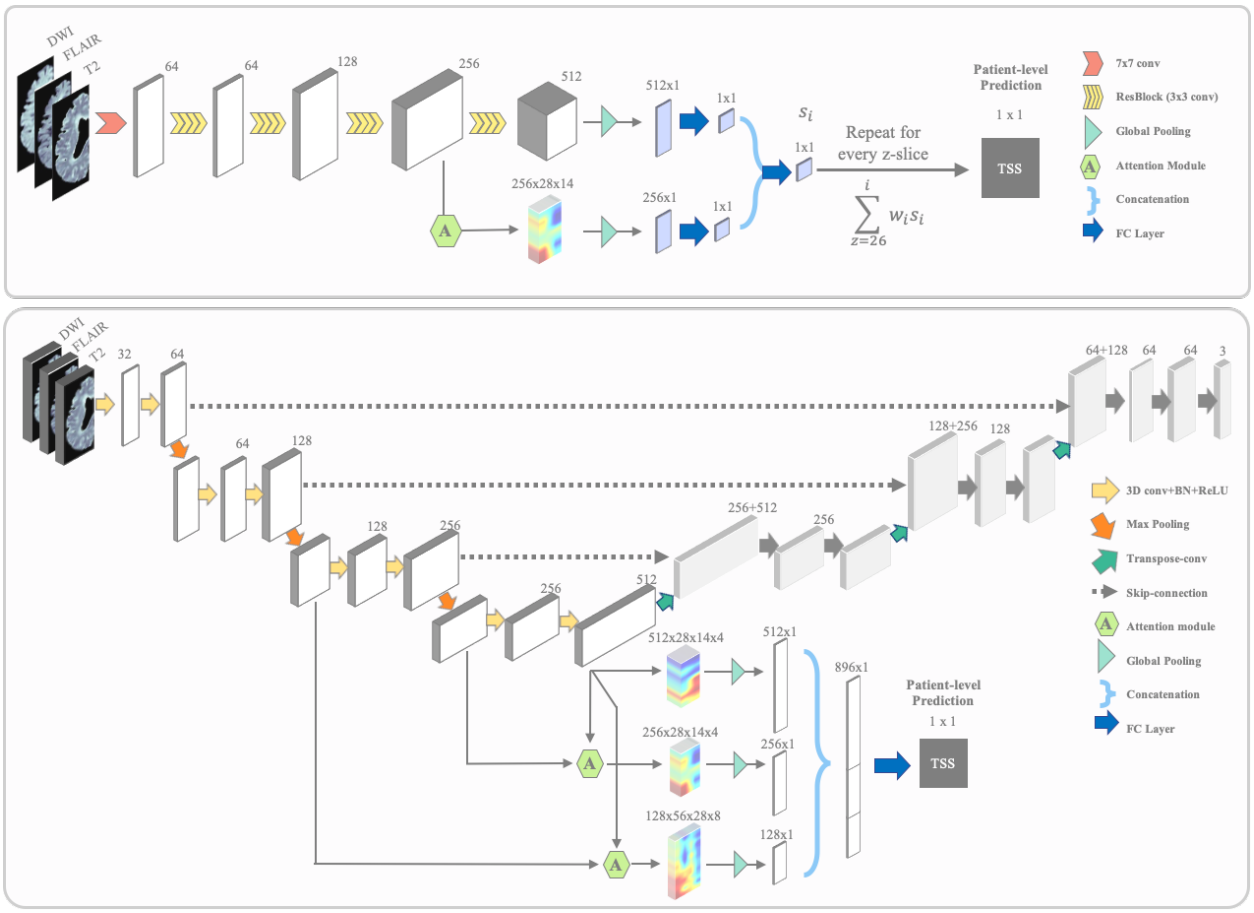


Figure 3.4: Architectures for 2D (top) and 3D (bottom) models. Our 2D Self-weighted Slice-wise Attention model took DWI b1000, T2w(b0), and FLAIR as a 3-channel input to a feature extraction backbone. Each slice of the brain was individually fed through four Resblocks of ResNet-18 to generate a 512x7x4 feature map, then pooled to a 512x1 feature vector [HZR16]. A soft attention module at the 256-channel convolutional layer was added to generate a 256x28x14 attention feature map and then pooled to a 256x1 feature vector. The feature map and attention feature map were aggregated for each slice with a learnable weighting factor for final classification. Our 3D model first used the entire structure of a 3D U-Net to train an initial weight using Models Genesis. Then volumetric DWI, T2w, and FLAIR were directly fed into the encoder part of the network. Two soft attention modules were added at 128 and 256-channel convolution layers. Feature maps from the original network and the two attention modules were pooled globally and concatenated for classification.

potential for over-fitting. To address these challenges, we first adapted a self-supervised learning approach, known as Models Genesis [ZZ19], to train a full 3D U-Net in order to generate initial weights for the stroke detection task. Using Models Genesis, we first modified the original images using non-linear transformation, local shuffling, in-painting, and out-painting and then trained the model to restore the original image, enabling the model to learn important high-level features in the original image. We then used the encoder component of the 3D U-Net network, along with two soft attention modules, to train this classification model to detect the stroke side and classify TSS. Figure 3.4 illustrates the 2D Self-weighted Slice-wise Attention Model structure and the 3D Attention Model structure. The Models Genesis and soft attention modules bolstered 3D model performance.

3.4.2 Training Schema

To train the models, each brain volume was split into hemispheres along the midsagittal plane on the registered volume. For each hemisphere, three imaging series, T2w, DWI, and FLAIR, were concatenated and input as channels with values normalized to a range of 0 to 1 and input dimension of 112x224x26. The right hemispheres were flipped on the vertical axis in order to spatially align with the left hemispheres for inputs. Our models used a multi-phase training regimen. The first phase consisted of stroke detection, where hemispheres were fed into the model separately and labeled as positive (1) if they had a stroke lesion in the hemisphere and as negative (0) if they had no stroke lesion in the hemisphere. The 2D model was trained from random initialization on this task. For our 3D model, initial weights were generated in a self-supervised fashion before the stroke side detection task for more rapid convergence. Once the model finished training, the first two convolutional layers/blocks were frozen. Specifically, for 2D models, in the ResNet-18 backbone, we froze the first 7x7 convolutional layer as well as the following two Resblocks, where the 7x7 convolutional layer is denoted conv1 and the two Resblocks each contains two 3x3 convolution layers denoted conv2_x from table 1 of [HZR16]. For 3D models, we froze the two layers in the downward path of the 3D U-Net backbone. As

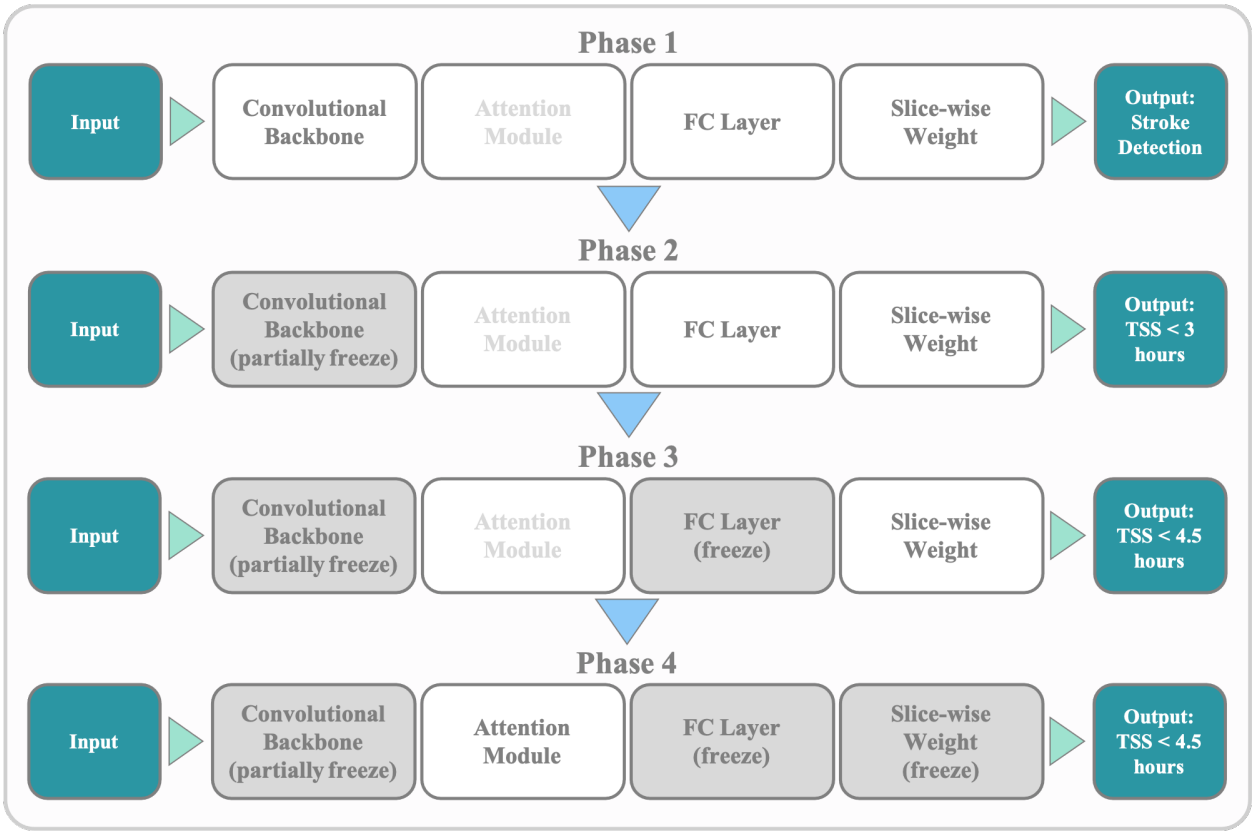


Figure 3.5: A summary of our training schema. Each phase utilized a unique classification label, as enumerated in the Outputs boxes for each phase. At the end of each training phase, the weights of certain components were frozen; these frozen weights were then initialized for the model at the start of the following phase.

described in [CAL16], each layer represents two 3x3x3 convolutions each followed by a ReLU, then a 2x2x2 max pooling with strides of two. This pre-trained network was then utilized in a second phase of training, whereupon only hemispheres with stroke lesions (positive cases in the stroke side detection task) were used as input. In the second phase, we froze early convolutional weights to refine later layers and trained our model on TSS < 3 hours, given the clinical correlation of DWI-FLAIR mismatch to this binarization. For the third phase, we used the pre-trained weights of the TSS < 3 hours model to train on the TSS < 4.5 hours task. The last phase of our training schema (Figure 3.5) involved fine-tuning the soft attention modules to further enhance performance. We compared this multi-phase training regimen to training on TSS labels from scratch, pretraining on natural images, and pretraining on external datasets of brain MRIs [CHC15, BSM19].

3.5 Experiment and Results

3.5.1 Evaluation Metrics

We trained the stroke detection algorithm for 100 epochs with early stopping, minimizing binary cross-entropy loss functions. All models were trained with the AdaBound optimizer [LXL19], which used bounds on a dynamic learning rate to transition smoothly from an adaptive method to the more traditional stochastic gradient descent. This approach allowed the model to maintain a higher rate of convergence in early training epochs. Hyperparameters were selected using a validation set during training. The batch size was 16 for the stroke detection task and 8 for the TSS classification tasks. The early stopping criteria was based on the validation AUC during training with a patience of 10. For Adabound, in the stroke side detection task, the initial learning rate was 0.0005 and the final learning rate was 0.01; in the TSS classification task, the initial learning rate was 0.00001 and the final learning rate was 0.001. The code was written in PyTorch, and experiments were run on an NVIDIA DGX-1. Our memory usage during training for the 2D models was 4GB VRAM with batch

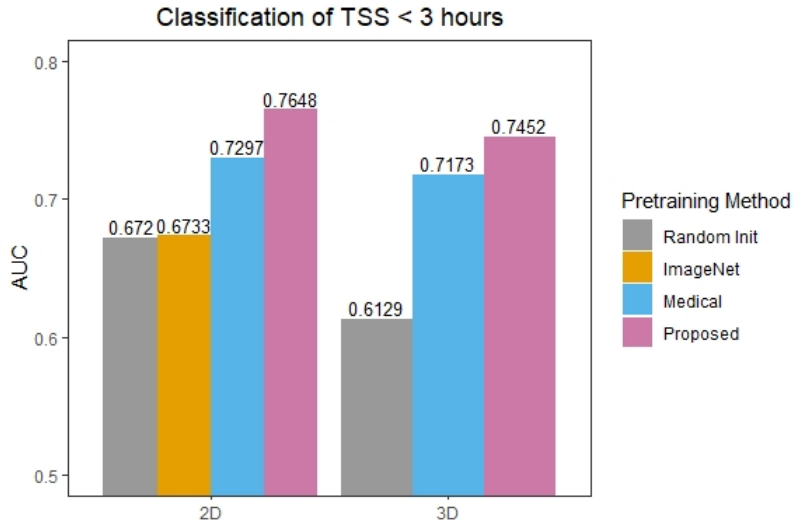


Figure 3.6: On second phase task TSS $<$ 3 hours, for 2D model, our proposed transfer learning approach has a 5.1% increase, whereas for the 3D model, there is a 8.3% increase in ROC-AUC score.

size 8 and 6GB VRAM with batch size 16; for the 3D models, memory usage was 7GB VRAM with batch size 8 and 12GB VRAM with batch size 16.

3.5.2 Results

The performance metrics for all of our training phases are summarized in Table 3.2. For stroke detection, the 2D and 3D architectures achieved ROC-AUC values of 0.8905 and 0.9460, respectively. This indicates that the models were able to reliably identify stroke at both the slice and volume level, which aligns with intensity differences usually observed for stroke lesions on DWI and FLAIR series. For the second training phase, classifying TSS $<$ 3 hours, our pretraining approach improved the performance of 2D model by 14.0% and our 3D model by 21.6% when compared to random initialization or to pretraining on natural images (2D model only). For both models, we also examined TSS classification performance with weights pre-trained on medical image datasets. We used models trained for brain tumor classification and segmentation to initialize our 2D and 3D models, respectively,

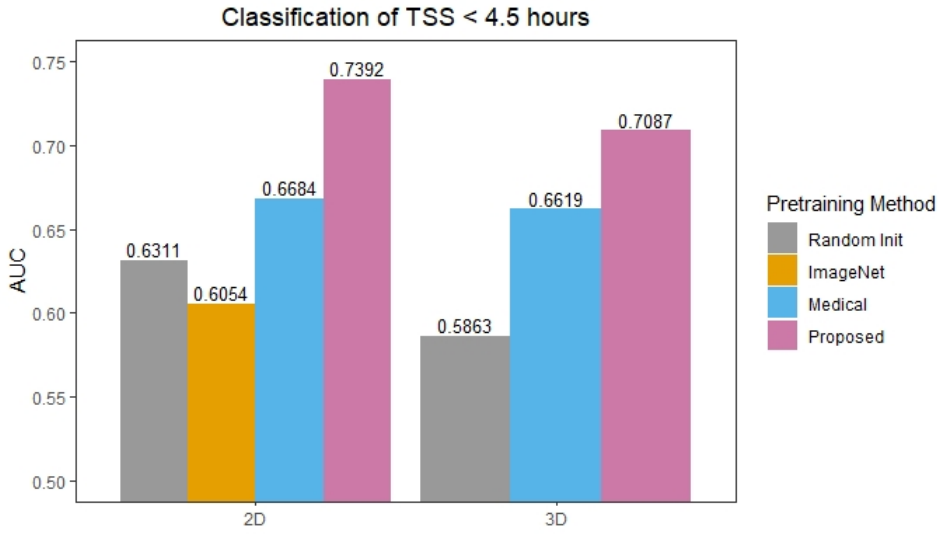


Figure 3.7: On third phase task TSS $<$ 4.5 hours, for 2D model, our proposed transfer learning approach has a 22.1% increase in AUC; for 3D model, there is a 20.9% increase

given that these tasks are in the same domain and use the same medical imaging modalities [CHC15, BSM19]. We froze the weights from earlier layers for both models, and we compared the effect of this pretraining to frozen weights learned from our stroke detection task. While performance improvement was observed using medical image pretraining, our pretraining approach was able to achieve higher performance for both models when compared to both natural image and domain-specific pretraining, with the 2D and 3D models achieving 76.48% and 74.52% increase in AUC, respectively.

In the third phase, we train the models to classify TSS $<$ 4.5 hours using weights from the second phase. As shown in Figure 3.7, both the 2D and 3D models improved classification performance by 22.1% and 20.9%. For the 2D model, pretraining on natural images reduced performance, which has been observed for other medical-image-specific tasks [RZK19]. As in Phase 2, We also show the results from ImageNet, Tumor detection, and segmentation weight transfer for comparison. As expected, due to the similarity of the dataset, the performance improvement is high, from AUC 0.6311 to 0.6684 and from 0.58 to 0.66 for the 2D and 3D models, respectively. However, the performance improvement (12.9% and 5.9%) is still

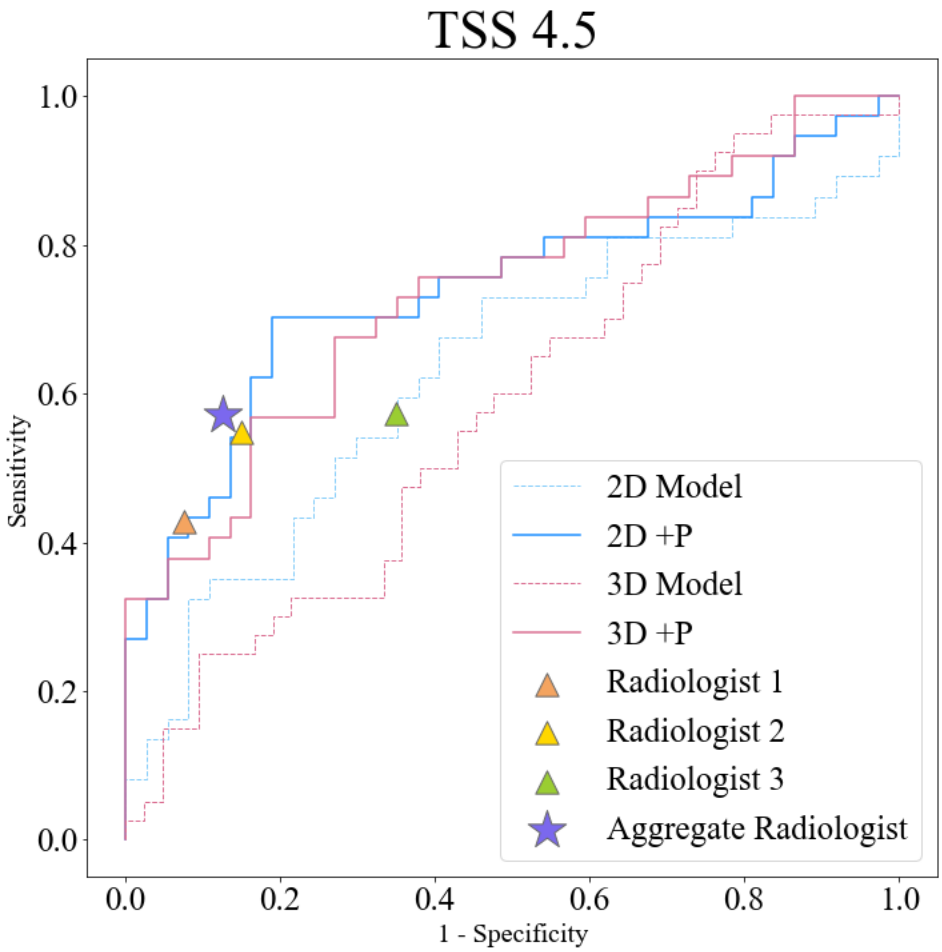


Figure 3.8: ROC curves for classifying TSS < 4.5 hours. +P = with pretraining.

lower than our proposed method (17.1% and 20.9% to AUC 0.7392 and 0.7087). For both tasks, the 2D model achieves higher performance than the 3D model, even with random initialization.

In the last of our proposed training phases, fine-tuning the attention modules yields improved performance for both the 2D and 3D models, though the improvement was more notable for the 3D model. The optimal ROC-AUC scores for classification of TSS < 4.5 hours are 0.7407 and 0.7370 for 2D and 3D respectively with 17.4% and 25.7% performance gain compared to training from scratch.

Stage	**Model**	**Weights**	**Sens.**	**Spec.**	**Acc.**	AUC
Phase 1	2D	Random	0.7347	0.9286	0.8316	0.8905
Stroke Detection	3D	Random	0.7732	0.9579	0.8646	0.9460
Phase 2	2D	Random	0.2444	0.9310	0.5135	0.6720
TSS $<$ 3 hrs		ImageNet	0.7879	0.5510	0.6463	0.6733
		Medical	0.6970	0.7142	0.7073	0.7297
		Phase 1	0.8222	0.6552	0.7568	**0.7648**
	3D	Random	0.7143	0.4848	0.6220	0.6129
		Medical	0.5952	0.7750	0.6829	0.7173
		Phase 1	0.8904	0.6000	**0.7724**	0.7452
Phase 3	2D	Random	0.2162	0.9189	0.5676	0.6311
TSS $<$ 4.5 hrs		ImageNet	0.8789	0.4285	0.6098	0.6054
		Medical	0.6666	0.6939	0.6829	0.6684
		Phase 2	0.5405	0.7838	0.6622	**0.7392**
	3D	Random	0.3750	0.6429	0.5122	0.5863
		Medical	0.8788	0.4489	0.6220	0.6619
		Phase 2	0.6279	0.7895	**0.7037**	0.7087
Phase 4	ML		0.6522	0.7143	0.6363	0.7174
TSS $<$ 4.5 hrs	2D	Phase 3	0.7027	0.8108	**0.7568**	**0.7407**
attention+fine-tune	3D	Phase 3	0.5405	0.8378	0.6892	0.7370
DWI-FLAIR	Rad 1		0.5476	0.8500	0.6951	
Mismatch	Rad 2		0.4286	0.9250	0.6707	
	Rad 3		0.5714	0.6500	0.6098	
	Agg Rad		0.5730	0.8750	0.7195	

Table 3.2: Performance metrics across tasks and architectures. Double lines separate models with different outputs. Sens = Sensitivity, Spec = Specificity, Acc = Accuracy, AUC = Receiver Operating Characteristic Area Under Curve, Rad = Radiologist, Agg Rad = Aggregate Radiologist.

For each model, we computed Youden's J statistic and reported the sensitivity, specificity, accuracy, and ROC-AUC score. We compared our model to the performance metrics of each radiologist's DWI-FLAIR mismatch assessments, which served as a proxy for TSS. We also compared our model to the previously-published model with the highest performance metrics by applying this model to our own dataset [LLH20]; these metrics are included in Table 3.2. Of note, the inter-reader agreement (Fleiss' kappa) was 0.46 among all three radiologists, which is typically regarded as a moderate level of agreement and aligns with previous findings of high variability among reader assessments. We also reported the ROC-AUC curves for each of our models in Figure 3.8.

We generated GradCAMs to visually assess model activation. To evaluate the utility of GradCAMs in a clinical context, an expert radiologist evaluated the overlap of the activation map and stroke lesion. The radiologist found that, for slices most representative of stroke lesion, 96% of cases evaluated had substantial overlap (>50%) between the lesion and activation, while the remainder of cases had moderate overlap. This indicates that Grad-CAM can qualitatively localize to stroke lesions when trained on the TSS tasks.

3.6 Discussion

Among the models tested, the pre-trained 2D model achieved the highest performance metrics with a sensitivity of 0.70 and a specificity of 0.81 in classifying TSS < 4.5 hours. Our model was more sensitive than the DWI-FLAIR assessments performed by the neuroradiologists, which we treated as a surrogate for determining a TSS < 4.5 hours. We also compared our model to the previously published state-of-the-art method. The threshold method implemented in [LLH20], which was used to create the ROI, was very stringent, in that only 221 of our original 422 patients had a large enough ROI from which features could be extracted. Thus, their performance metrics represent a subset of our larger dataset. We also tested our model performance on this subset and achieved a ROC-AUC of 0.76. Nevertheless, on the

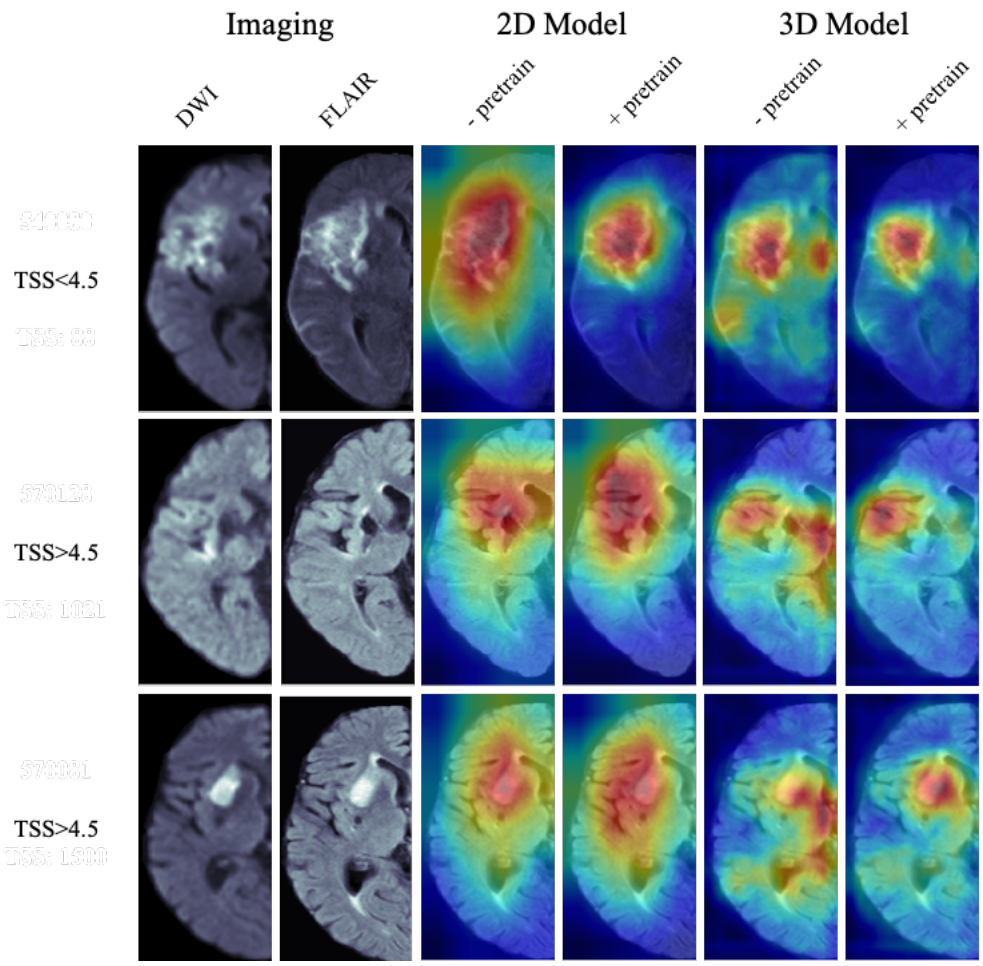


Figure 3.9: Grad-CAM visualizations of the penultimate convolutional layer for 2D and 3D models, both from scratch and with pretraining.

entire dataset, the optimal 2D model with pretraining was able to outperform the previous model. From a clinical perspective, these results indicate that our model may be able to correctly identify more patients within the 4.5-hour window and therefore eligible to receive thrombolytic therapy when compared to both DWI-FLAIR mismatch assessment and the threshold-based Machine Learning (ML) method. There are many tasks within the medical image domain to which our proposed task-adaptive pretraining schema can be applied. For example, this schema could be used for brain tumor classification, where brain tumor detection is the pretraining task.

The optimal 2D model has a ROC-AUC comparable to that of the 3D model; however, the sensitivity (0.54) and specificity (0.84) of the 3D model are less balanced, indicating that while the rate of true negatives is high, there are less true positives identified by that model. In total, our model metrics illustrate that the progressive pretraining schema enhances performance for our task considerably, for both our proposed 2D and 3D architecture. For both models, attention modules enhance performance. The use of GradCAM for our models highlights regions of the brain that impact decisions, as illustrated in Figure 3.9. The GradCAMs illustrate that the pre-trained model is able to more precisely localize to the stroke infarct and highlight other regions outside of the infarct that may inform this classification task.

Our model performance metrics are comparable to previous approaches in TSS classification. However, this study has a few factors that increase its potential clinical applicability. The patients in our dataset comprise a wider range of stroke locations and other clinical demographics than in previously assessed datasets. Additionally, our model leverages the entire brain hemisphere, which may contain more relevant information among this broader patient cohort. This has the potential to reduce bias in our model, and the convolutional architecture, allows this information to be incorporated into decision-making.

That said, DL generally requires a high volume of data. While many medical image-related tasks have used DL with a comparable amount of patient data used here, a higher volume of data would greatly enhance the model performance. This model only uses diffusion-

based imaging, as these are the images used in current clinical practice. Incorporating perfusion-based imaging and its derivatives such as perfusion maps may better inform TSS. There is a substantial body of work using perfusion imaging parameters for stroke outcomes [Sca13, dTP17, HSZ19, LLH20]. Based on our examination, registration quality was not affected by the ischemic lesion in the T2w images, as the lesion was not apparent in the T2w sequence. While this type of registration failure was not a concern in our dataset, it could possibly affect other neuroimaging studies. Finally, the use of clock time as a label for TSS may not fully encompass the physiology underlying ischemia in the brain; for example, the cerebral collateral flow may compensate for a hypoperfused area within the brain and reduce the amount of ischemia that tissue is experiencing during a stroke [Ban11], which may be the biological reason for DWI-FLAIR mismatch.

3.7 Summary

This approach uses 2D and 3D CNN models to classify TSS for 422 patients and compares model performances to DWI-FLAIR mismatch readings performed by three expert neuroradiologists. We demonstrate that our 2D model outperforms the 3D model when classifying TSS < 4.5 hours, which is the current clinical guideline. We show that pretraining the model on stroke detection, then refining the model on TSS classification yields better performance than training on TSS classification labels alone; the incorporation of soft attention modules also enhances the performance of both the 2D and 3D when compared to CNNs without them. By visualizing network gradients via Grad-CAM, we show that our pre-trained models localize to stroke infarcts and surrounding regions. We demonstrate that both our 2D and 3D model is able to generalize to an inclusive dataset comprising multiple types of ischemic stroke and that this model may be able to inform TSS for patients with unknown symptom onset.

The results comparison demonstrated our proposed DL approach in Chapter 3 exceeds

the previous state-of-the-art ML method in all settings, showing a more robust automated algorithm to determine stroke onset time using pretreatment MRI.

3.8 Appendix

3.8.1 Additional Experiments Results

We further evaluate our 2D model on an extended dataset from Asan Medical Center [LLH20] and compare it with their proposed ML approach. We also conduct a DWI-FLAIR mismatch reader study on both internal and external datasets for comparison.

3.8.1.1 Data

The internal dataset is comprised of a similar cohort with slightly different inclusion criteria to align with the external dataset. Individuals were included in the cohorts based on the following inclusion criteria: (1) diagnosis with AIS, (2) received pretreatment MRI protocol with DWI, FLAIR, and apparent diffusion coefficient (ADC) series without motion degradation, and (3) known TSS within 24 hours of image acquisition. The internal cohort comprised 417 patients treated from 2011 to 2019. The second dataset, published by Lee et al[LLH20] totaled 355 patients, with more extensive exclusion criteria than previously described. To ensure consistency across both datasets, images were subjected to a preprocessing pipeline [ZPN21b]. Demographics are shown in 3.3.

3.8.1.2 Experiments and Results

The comparing ML approach uses a thresholding approach to generate the infarct Region of interest (ROI), then extract the radiomics features from DWI, FLAIR, ADC, and FLAIR-ADC ratio maps and select a subset of the generated radiomics features using statistical testing. Selected features are then fed into Random Forest, Support Vector Machine, or Logistic Regression for classification. The comparison results of ML and DL are shown in 3.4.

	Internal		External	
	Train (n = 343)	Test (n = 74)	Train (n = 299)	Test (n = 56)
Age (years)	70 (55-80)	68 (57-79)	63 (55-73)	67 (55-71)
Female	176 (52%)	46 (56%)	86 (34%)	20 (36%)
NIHSS	8 (4 - 16)	6.5 (2 - 18)	4 (2-10)	5 (2-12)
Onset to MRI (min)	210 (105-683)	230 (107-661)	270 (152-712)	240 (142-448)
In 4.5 hours onset (%)	185 (54%)	37 (50%)	153 (58%)	24 (43%)

Table 3.3: Internal and external patient cohort demographics. Numbers are n (%) or median (interquartile ranges); NIHSS, National Institutes of Health Stroke Scale.

Method	**Train set**	**Test set**	**AUC**	**Acc.**	**Sens.**	Spec.
Deep Learning	Internal	Internal	.768 ± .03	.726 ± .02	.712 ± .08	.741 ± .09
		External	.737 ± .03	.724 ± .04	.757 ± .04	.679 ± .07
	External	Internal	.732 ± .02	.707 ± .03	.716 ± .09	.687 ± .08
		External	.772 ± .02	.767 ± .03	.850 ± .08	.648 ± .09
	Both	Internal	.840 ± .03	.789 ± .04	.777 ± .06	.802 ± .07
		External	.814 ± .01	.800 ± .04	.850 ± .08	.727 ± .08
Deep Learning	Internal	Internal	.730 ± .07	.675 ± .07	.405 ± .07	.811 ± .08
		External	.680 ± .15	.653 ± .10	.714 ± .15	.500 ± .13
	External	Internal	.698 ± .08	.625 ± .09	.297 ± .08	.865 ± .10
		External	.780 ± .05	.735 ± .05	.657 ± .05	.800 ± .08
	Both	Internal	.783 ± .03	.750 ± .04	.405 ± .03	.892 ± .03
		External	.795 ± .03	.735 ± .03	.686 ± .03	.750 ± .04

Table 3.4: Performance metrics for Deep Learning (DL) and Machine Learning (ML). Models trained on internal, external, or both and tested on the internal and external test sets. Sens = Sensitivity, Spec = Specificity, Acc = Accuracy, AUC = Receiver Operating Characteristic Area Under Curve

CHAPTER 4

Predicting Thrombectomy Outcomes Using Machine Learning and Deep Learning Approaches

4.1 Introduction

Self-attention mechanism has been shown to help the CNN model to focus better on regions of interest without explicit segmentation mask supervision. Self-attention-enhanced CNN can reduce data usage and increase the speed of convergence during training. In this chapter, we aim to predict thrombectomy outcomes using CNN-based models specifically tailored to small datasets and thick-sliced stroke imaging, which incorporates CNN-transformer hybrid architecture and slice cross-attention aggregation. First, we explored the capability of pre-treatment MRI in predicting mTICI scores using radiomics features and Machine Learning (ML) algorithms [ZPN21a]. After the exploratory study confirming correlations, we further applied deep learning (DL) models to predict mTICI scores using pretreatment CT images [ZPY23]. Finally, we modified the DL model to predict the first-pass effect.

4.2 Overview

Mechanical Thrombectomy (MTB), also known as Endovascular Thrombectomy (EVT), stands as the primary treatment for patients experiencing clots in large blood vessels. This procedure involves the surgical removal of a blood clot from an artery, with the objective of achieving recanalization, or in simpler terms, restoring blood flow. The success of an EVT

procedure is determined by the extent to which blood flow is reinstated, ideally to a full or near-full extent, to the brain region affected by the stroke. To gauge the level of recanalization achieved, patients undergo an evaluation post-treatment, receiving a score based on the modified Treatment in Cerebral Ischemia (mTICI) scale [Tom07, DCB17, FKK13]. This post-treatment score is clinically significant, as it has been shown that favorable scores, i.e., mTICI 2c or greater, are associated with better long term clinical outcomes [CBL17]. Unfavorable scores (mTICI less than 2c) indicate that the treatment did not effectively clear the blood vessel. Clinical trials have illustrated that patients who experience significant and/or full recanalization of the blood vessel typically experience better outcomes, particularly if recanalization is achieved on the first attempt – known as the first pass effect (FPE) [LFH19, BRC17, DPG20, FBF21]. Imaging has been identified as one modality to illustrate patient physiology that could influence the likelihood of a successful EVT procedure. Predicting the final mTICI score and FPE prior to a procedure can provide doctors with more information when considering treatment options.

The available evidence indicates that despite similar clinical history, stroke characteristics, and procedural factors, patients tend to experience different outcomes when it comes to recanalization. In an effort to better understand the factors that contribute to a patient's likelihood of successful recanalization, several early studies have been conducted, many of which have relied on non-invasive imaging methods to identify clinical correlates. Similar to assessments for thrombolytic therapy, the amount of time that has elapsed since the onset of the stroke has been shown to be positively associated with long-term clinical outcomes following EVT. Moreover, the identification of the penumbra region through the use of MR or CT imaging has proven to be useful in determining treatment outcomes. In addition to these factors, the presence of compensatory flow from collateral circulation, known as collateral flow, has also been found to have a strong correlation with prognosis following EVT. [Ban11, PSG19, SBM11]. DL has been shown to leverage the amount of detail in images to improve prediction accuracy in neuroimaging, such as stroke or glioblas-

toma [YYZ, TZY22, HSZ19, ZPN21b]. Current literature presents models that perform semi-automated prediction of mTICI score based on pre-treatment CT imaging that relies on manual segmentation of the clot by an expert neuroradiologist which does not fit the urgency of EVT treatment under current guidelines [HBR20, QKN19, HRO19].

Our exploratory studies around EVT treatment outcome predictions using pretreatment imaging are as follows: We first propose a fully automated ML model to classify mTICI scores using pretreatment MRI [ZPN21a]. Second, we designed a hybrid transformer DL model to predict mTICI using pretreatment CT and CTA [ZPN21a]. Last, we improve the design of the DL model to predict FPE using pretreatment MRI and CT to evaluate the predictive capability of both modalities. In the first two studies, both ML and DL approaches on CT and MRI showed promising performance. In this chapter, we introduce in detail our efforts in predicting FPE using both MRI and CT. We hypothesize that DL may extract useful information from pretreatment MR and CT that are correlated with FPE. We report performance metrics for two cohorts of patients: those who underwent CT and MRI before treatment, respectively. We design the framework specifically tailored to the small sample size and the thick slice nature of pretreatment stroke imaging. We incorporate contrastive learning to leverage a larger imaging dataset of AIS patients that triage to different treatment avenues and/or have missing clinical information to pretrain the model, thus helping the model to better generalize on a small dataset.

4.3 Method

4.3.1 Dataset

The cohort used for this study was retrospectively collected from UCLA Ronald Reagan Medical Center from 2014-2021. AIS patients were included if they were diagnosed with a large-vessel occlusion (LVO), had an MRI or CT acquired upon admission under stroke protocol, and received EVT treatment. Exclusion criteria were as follows: the presence of

significant hemorrhage and failed image registration, and low image quality such as motion blur or spike. As part of the EVT protocol at UCLA, mTICI is assessed during the procedure after each clot retrieval pass. This study defined successful recanalization as an mTICI of 2b, 2c, or 3. Baseline features such as age, sex, NIHSS at admission, and time since stroke, were compared between patients who did or did not experience FPE using the chi-square test, student's t-test, and Wilcoxon's rank sum test as appropriate. The cohort's clinical, imaging, and procedural characteristics are listed in Table 1. All statistical analysis was performed using R software 4.1.3(https://www.r-project.org).

4.3.2 MR Acquisition and Preprococessing

Patient MR imaging was acquired on 1.5T and 3T echo-planar MR scanners with 12-channel head coils (Siemens, Germany). In the stroke MRI brain imaging admission protocol, the diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) sequences were acquired using the following parameters: DWI: TR 4000-9000, TE 78-122ms, corresponding pixel dimensions 0.859x0.859x6.000 to 1.850x1.850x6.500 mm; FLAIR: TR 8000-9000ms, TE 88-134ms, corresponding pixel dimensions 0.688x0.688x6.000 to 0.938x 0.938x6.500 mm. Apparent diffusion coefficient (ADC) maps were calculated from DWI b0 and DWI b1000 using the following formula:

$$ADC = -ln\frac{Sb1000/Sb0}{1000} \tag{4.1}$$

Where $Sb1000$ and $Sb0$ are the intensity values of DWI b1000 and DWI b0 images. From MRI, the series used included DWI, FLAIR, and ADC sequences, Automated preprocessing steps described in [ZPN21b] is performed to segment vascular regions for stroke. Briefly, all sequences are subjected to N4-bias field correction using the ANTs library [ATS09], intensity normalization, and histogram matching. Finally, registration to MNI-space enabled the use of a vascular territory atlas for stroke region localization. Sample MRI original images and the processed inputs are shown in Figure 4.1.

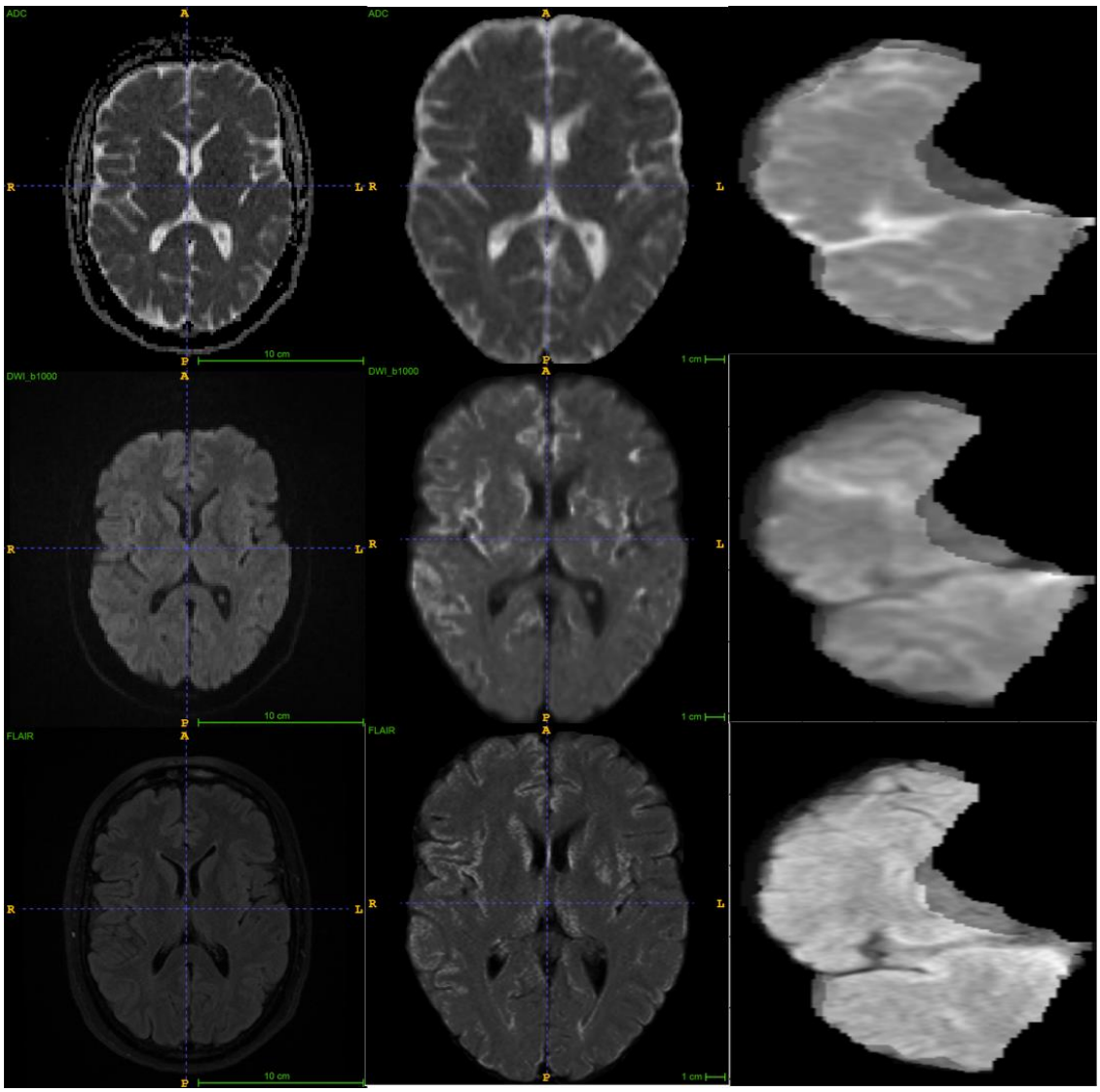


Figure 4.1: Sample DWI and FLAIR images. Left are original images, middle are registered images, and right are mapped regions for input

4.3.3 CT Acquisition and Preprococessing

Two CT scanners, a Lightspeed VCT (GE Health Care, Milwaukee, USA) and a SOMATOM Definition (Siemens, Forchheim, Germany), were used for CT imaging. After administering 50 mL of contrast agent intravenously at 5 mL/second, a single-phase CT-angiography (CTA) was obtained (120 kV, 120 reference mAs, 0.3 second rotation time, 0.6 pitch, effective dose of about 3 mSv). Following intravenous injection of contrast agent, totaling 50 mL at a rate of 5 mL/second, CTP included 30 successive spiral acquisitions (80 kV, 150 mA, effective dose = 3.3mSv, 100 mm in the z-axis) in a total of 60s acquisition. Saline was used after each contrast agent injection, with 30mL being used for each injection. Both non-contrast CT (NCCT) and CTA image series were included as inputs for the imaging-based models. The preprocessing protocol for CT images included field-of-view removal, skull stripping, and registration to MNI space. Sample CT original images and the processed inputs are shown in Figure 4.2.

4.3.4 Deep Learning Model Architecture

The proposed DL model is an end-to-end hybrid neural network consisting of both convolutional and transformer attention components, namely the multi-sequence neighborhood transformer model (MNT-DL). MNT-DL is a hybrid transformer architecture that incorporates modifications and enhancements to the widely used ResNet backbone. Each convolutional block is modeled after ResNet residual blocks, consisting of the following sequence: convolutional kernel, batch normalization, rectified linear unit activation function, second convolutional kernel, and second batch normalization. The initial component is a global feature extractor that utilizes residual convolutional blocks to extract relevant features from every slice. Subsequently, the extracted slice-level features are passed onto local networks that are designed to learn representations of neighboring slices, while also sharing weights during the training process. In the local network, a self-attention module is utilized to identify the

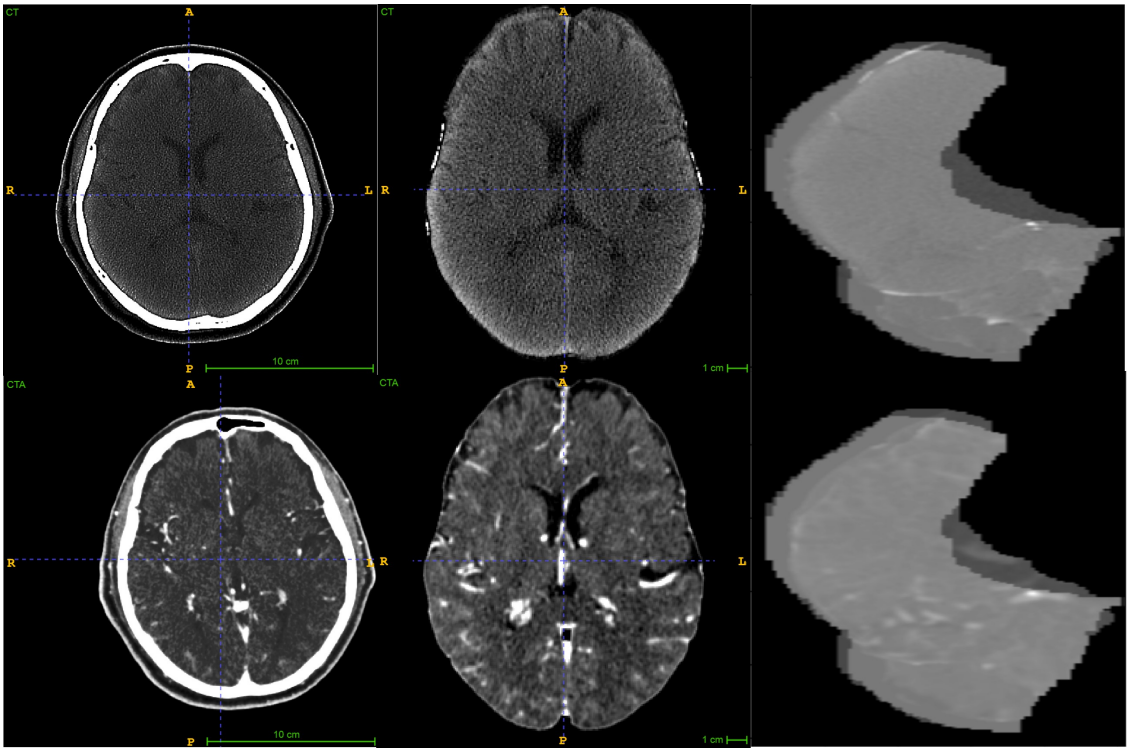


Figure 4.2: Sample NCCT and CTA images. Left are original images, middle are registered images, and right are mapped regions for input

most significant regions within each slice. The self-attention module uses a 1x1 convolution on the intermediate features to generate single-head attention for each image patch, computing attention with respect to all other patches. The significant regions identified in each slice using the self-attention module are then combined through matrix multiplication and SoftMax activation. The transformer self-attention modules are integrated into the network easily without incurring significant computational overhead.

After passing through the local networks, the outputs are fed into the volumetric classifier, which comprises two modules. The first module is a cross-attention module. The low-level features extracted from each slice are fed into the cross-attention module, which employs multi-head attention operations to generate slice-level importance. Similar to other attention modules, multi-head attention involves a linear layer that generates attention across multiple scales of the image volume. The attention operations are fused using cross-attention, where the features from each scale are exchanged through layer normalization and residual connection. By using this module in the network, the model is able to assign greater weight to the slices that are more important for the final prediction, while incurring only limited computational complexity. The output of this attention module, along with the output from the local networks, is then fed into a linear layer that serves as the final classifier, producing the volume-level prediction. DWI-FLAIR-ADC sequences or NCCT-CTA sequences are used as channels to input into MNT-DL for MR or CT input. Single-sequence inputs are also used to develop corresponding models (ADC-DL, DWI-DL, FLAIR-DL, NCCT-DL, CTA-DL) to for ablation studies where the single sequences are stacked to fit into the same channel requirement for corresponding models.

4.3.5 Contrastive Self Supervised Learning

Although we use multiple model designs tailored for small sample size in DL training, the DL training still limited by the labeled data for MRI and CT scans. Therefore, we adopt a contrastive self-supervised learning (SSL) approach called SimSiam [CKN20] to our proposed

model. SimSiam does not require a large batch size, negative sample pairs, and a momentum encoder. Under this approach, we facilitate more imaging data from our institutional stroke registry that do not meet this study's inclusion criteria, and further improve the model's performance. The model architecture and self-supervised learning framework are shown in Figure 4.3.

4.3.6 Loss Function

The loss function used in this work was based on binary cross-entropy, defined as:

$$Total\ Loss = L_{fusion} + \gamma * (L_{subnet_1} + L_{subnet_2} + L_{subnet_3} + L_{subnet_4} + L_{subnet_5}) \tag{4.2}$$

where L is binary cross-entropy loss. The fusion loss, L_{fusion}, denotes the loss of the final output of the global network. In addition, the loss is computed for the intermediate outputs of each local network L_{subnet_x}. The losses L_{subnet_1}, L_{subnet_2}, L_{subnet_3}, L_{subnet_4}, and L_{subnet_5} are added together and combined with L_{fusion} using weighting factor γ. In this study, the weighting factor was set at 0.5 to give equal weights between the final output loss and the sum of local network losses.

4.3.7 Training and Evaluation

Models were evaluated to predict a binarized FPE label for each patient. A patient was considered positive if they had an mTICI score of 2b, 2c, or 3 after one pass of clot retrieval. Patients that achieved recanalization in several attempts, or who did not achieve successful recanalization eventually, were negative. The MRI and CT cohorts were divided into retrospective development and prospective evaluation sets if they underwent EVT before or after the year 2020. Five-fold cross-validation was used for development. In each fold, the model was trained for 100 epochs with early stopping using the AdamW optimizer. The learning rate was set to 0.0005 and the weight decay was set to 0.05. The training was implemented using Pytorch 2.0 on an NVIDIA DGX-2. Following the development and

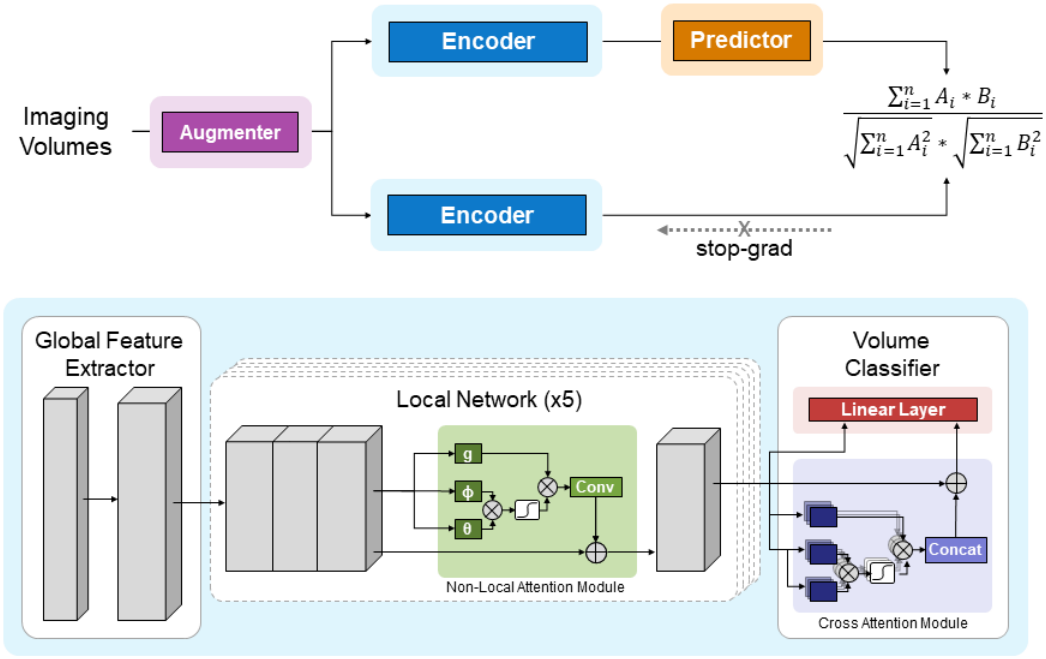


Figure 4.3: FPE prediction framework. The top is the self-supervised learning approach, the bottom is the model architecture

hyperparameter tuning, algorithms were evaluated on the corresponding prospective evaluation set. Receiving-operator characteristic area-under-the-curves (ROC-AUC) were reported accordingly. Sensitivity, specificity, and accuracy were calculated using Youden's J statistics from the ROC curve [You50]. All metrics were reported as mean±standard deviation on the evaluation set for each cohort that reflects the model performance across each fold.

4.4 Results

4.4.1 Patient Characteristics

This cohort included 408 patients who met the criteria; of these, 76 patients were excluded due to missing image series (52) or degraded image quality preventing preprocessing (24). The patient inclusion workflow diagram is shown in Figure 4.4. From this final cohort of 332 patients, 152 underwent MRI, and 180 underwent CT before EVT. The cohort had an average age of 71.49±15.94 years and was 54.22% female. Of this cohort, 80 patients experienced a stroke within 24 hours of the last-known well time but had an indeterminable onset time. Among patients with known onset time, 168 (50.60%) received imaging within the 4.5-hour window and 185 (55.72%) underwent contrast MRI or CT within 6 hours. Median NIHSS upon admission was 16 (IQR 10-20). Prior to EVT, 96 patients (28.92%) received intravenous thrombolytic therapy. Additional clinical variables as well as differences between the MRI and CT cohorts are summarized in Table 4.1. For the Self-supervised pretraining stage, we collected 599 MRI and 475 CT images from the UCLA stroke registry that meet image sequences and quality requirements for the preprocessing steps in our study but do not qualify for the EVT study due to different treatment triage, missing basic clinical information, etc.

4.4.2 Model Performance

The 5-fold cross-validation performance of the DL models on MRI was summarized in Table 2. The ROC-AUC of MNT-DL was higher than those of any single-sequence models (ADC-DL, DWI-DL, FLAIR-DL), achieving a mean ROC-AUC of 0.7505. Adding SSL further improved the ROC-AUC to 0.8443. Similarly, as shown in Table 3, MNT-DL for CT images achieved a ROC-AUC of 0.7801, higher than both NCCT and CTA single-sequence models (NCCT-DL and CTA-DL). SSL further improve the ROC-AUC of MNT-DL to 0.8719. The performance of the DL models on MRI and CT for both prospective test sets was summarized

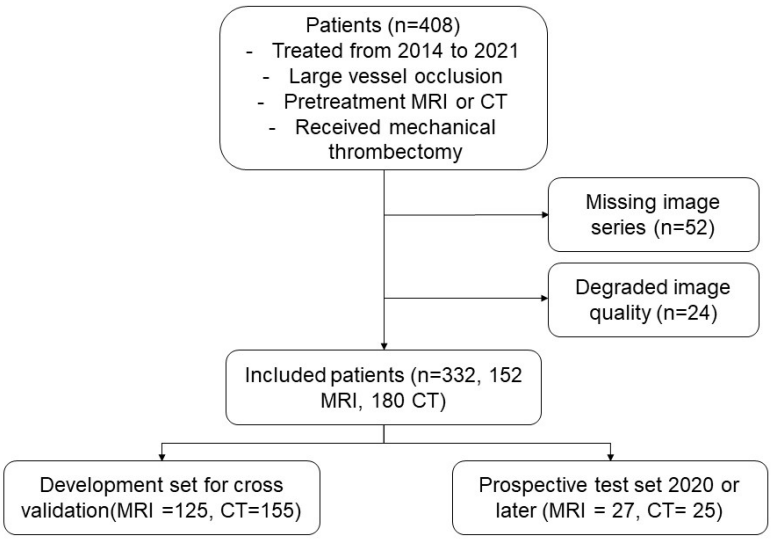


Figure 4.4: Inclusion Criteria for patient cohort

Parameters	Measure	Total (N=332)	MRI (N=152)	CT (N=180)	p-value
Age (years)	Mean±SD	71.49±15.94	70.72±16.11	72.13±15.77	0.4237
Female	N (%)	180 (54.22%)	86 (56.58%)	96 (53.33%)	0.6303
NIHSS	Median (IQR)	16 (10-20)	15 (8 - 19)	16 (11 - 21)	0.0496
Received tPA	N (%)	96 (28.92%)	35 (23.03%)	61 (33.89%)	<0.0001
Stroke Onset Time					0.6013
Onset Time (min)	Median (IQR)		167 (123-255)	113 (83-191)	–
Unknown Onset	N (%)	80 (24.10%)	42 (27.63%)	38 (21.11%)	–
Onset < 4.5 hours	N (%)	168 (50.60%)	71 (46.71%)	97 (53.89%)	–
Onset < 6 hours	N (%)	185 (55.72%)	77 (50.66%)	108 (60.00%)	–
Thrombectomy Outcome					0.5181
Unsuccessful	N (%)	59 (17.78%)	28 (18.42%)	31 (17.22%)	–
mTICI 0|1|2a	N|N|N	20|4|34	11|2|15	9|2|20	–
Successful, 2+ Passes	N(%)	133 (40.06%)	57 (37.50%)	76 (42.22%)	–
mTICI 2b|2c|3	N|N|N	73|31|25	37|12|8	37|20|19	–
Successful, First Pass	N (%)	140 (42.17%)	67 (44.08%)	73 (40.56%)	–
mTICI 2b|2c|3	N|N|N	59|34|43	31|14|22	31|21|21	–

Table 4.1: Demographics of patients included in model development. N, number of patients; SD, standard deviation, IQR, interquartile range; NIHSS, National Institutes of Health stroke scale; mTICI, modified treatment in cerebral infarction score; tPA, intravenous thrombolysis.

in Table 4. When applied to the MRI series, the DL model achieved an average ROC-AUC of 0.7967, with an accuracy of 0.7774 on the prospective test set. The ROC curves are shown in Figure 4.5. The model outperformed the previous method, notably achieving near-perfect specificity across experimental replicates while maintaining high sensitivity. In the prospective CT evaluation set, the DL method performed similarly, yielding a mean ROC-AUC of 0.8051 and an accuracy of 0.8080. Compared to the literature, this model achieved slightly lower average accuracy, though with a substantially narrower confidence interval. While the accuracy was slightly lower, the model achieved a more balanced sensitivity and specificity of 0.8615 and 0.7500, respectively, compared to the previous model that achieved high specificity at the expense of very low sensitivity.

Model	ROC-AUC	Accuracy	Sensitivity	Specificity
ADC-DL	0.7127 (0.0492)	0.7417 (0.0995)	0.8942 (0.1138)	0.6263 (0.1324)
DWI-DL	0.6887 (0.0405)	0.7083 (0.0659)	0.7058 (0.1084)	0.7715 (0.1544)
FLAIR-DL	0.6957 (0.5011)	0.7083 (0.0833)	0.7497 (0.2163)	0.7691 (0.1913)
MNT-DL	0.7505 (0.0438)	0.7875 (0.0342)	0.7326 (0.1717)	0.8366 (0.1363)
NT-DL+SSL	0.8506 (0.0712)	0.8625 (0.0280)	0.9350 (0.0929)	0.8057 (0.0944)

Table 4.2: Ablation study on MRI cross-validation folds

Model	ROC-AUC	Accuracy	Sensitivity	Specificity
NCCT-DL	0.7404 (0.0560)	0.7813 (0.0221)	0.7882 (0.1377)	0.7629 (0.0819)
CTA-DL	0.7385 (0.0535)	0.7812 (0.0442)	0.7823 (0.1080)	0.8026 (0.1246)
MNT-DL	0.7801 (0.0320)	0.7979 (0.0592)	0.7923 (0.1303)	0.8066 (0.0947)
NT-DL+SSL	0.8719 (0.0831)	0.8688 (0.0640)	0.9381 (0.0852)	0.8058 (0.1202)

Table 4.3: Ablation study on CT cross-validation folds

Model	ROC-AUC	Accuracy	Sensitivity	Specificity
MRI	0.7967 (0.0335)	0.7774 (0.0367)	0.7286 (0.1849)	0.8462 (0.1216)
CT	0.8051 (0.0377)	0.8080 (0.0299)	0.8615 (0.1131)	0.7500 (0.1054)

Table 4.4: Deep learning model performance on prospective MRI and CT test set

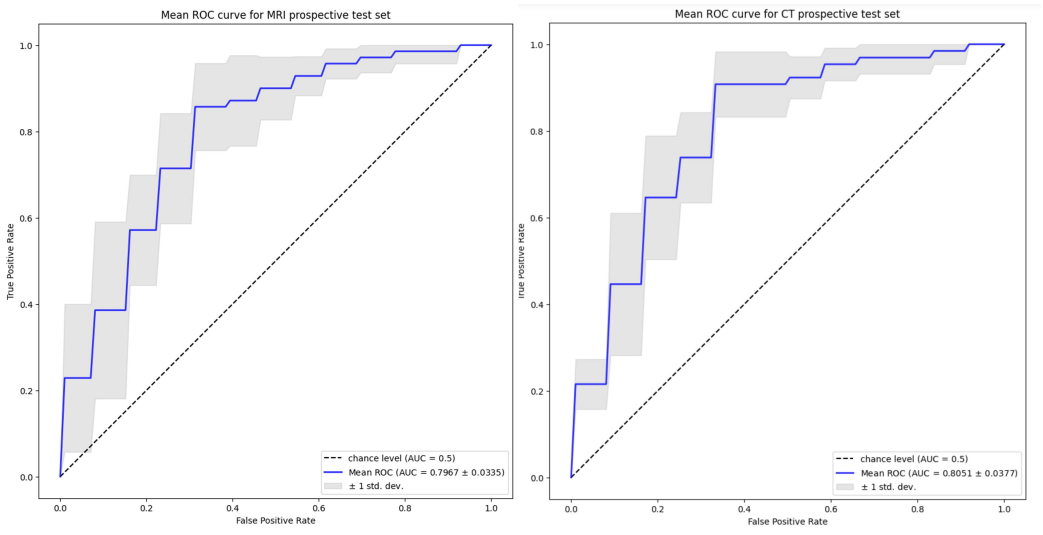


Figure 4.5: ROC curves both MR and CT test performance

4.5 Discussion

The phenomenon known as the First Pass Effect (FPE) has been demonstrably linked to improved patient outcomes in cases of Acute Ischemic Stroke (AIS). For more precise and effective planning of the Endovascular Thrombectomy (EVT) strategy, it's crucial to establish a reliable predictive association between the initial imaging performed before treatment and the likelihood of FPE. Our study aims to explore the potential of such pretreatment imaging in accurately predicting the occurrence of FPE during EVT procedures.

This study, to the best of our knowledge, introduces the first algorithm designed to predict the First Pass Effect (FPE) using pretreatment Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) scans of patients. Vital insights extracted from standard pre-treatment diffusion MR images and CT scans have shown a direct correlation with Endovascular Thrombectomy (EVT) recanalization, thus uncovering an innovative avenue for research into pre-treatment imaging and thrombectomy outcomes. Our usage of Deep Learning (DL) algorithms offers several advantages over traditional methods. Our methodology eliminates the necessity for manual clot segmentation, which is typically a time-intensive process that may inadvertently delay crucial treatment. We employ automated preprocessing, registration, and region map masking to streamline the input image region of interest. Consequently, our model autonomously identifies pertinent features from the input images, bypassing the need for manual intervention. Moreover, our innovative application of contrastive self-supervised learning underscores the effectiveness of Self-Supervised Learning (SSL) in scenarios where training data availability is limited. This serves as valuable evidence for utilizing similar approaches in medical imaging training. Finally, our models display high accuracy in predicting FPE without dependency on advanced imaging techniques like perfusion imaging. The latter, which involves a long acquisition time and may not be readily available in all stroke triage settings, is effectively sidestepped by our models. Thus, our approach is not only accurate but also versatile, promising swift and practical

solutions in diverse healthcare settings.

"Numerous studies have embarked on the mission to forecast recanalization for Acute Ischemic Stroke (AIS) patients. Utilizing just clinical variables has demonstrated only moderate success in this endeavor [AVC21, VMS21]. Slightly superior performance has been achieved by employing handcrafted or statistical features derived from manually segmented Regions of Interest (ROI). Yet, these methods still leave ample room for further enhancement in accuracy [MKC21, RBT14, SDS20, GGB20]. One study used radiomics features from manually segmented regions and ML models to predict FPE from CT images, achieving high specificity but low sensitivity [HBR20]. Contrastingly, our proposed methodology eliminates the need for manual segmentation and strikes a balanced sensitivity-specificity trade-off while maintaining competitive accuracy levels. The Deep Learning (DL) algorithm we developed was assessed on cohorts who underwent either Magnetic Resonance (MR) or Computed Tomography (CT) imaging prior to treatment. This evaluation illustrates that both these imaging modalities harbor substantial and valuable information related to the First Pass Effect (FPE). This study does present several limitations that need to be considered. Firstly, the potential for treatment bias is inherent in our model, as the cohort solely consisted of patients who had undergone Endovascular Thrombectomy (EVT) due to the study design. This could skew the outcomes in favor of EVT-receiving patients. Secondly, another layer of bias stems from the assignment of the modified Thrombolysis in Cerebral Infarction (mTICI) score. Given that a single neurointerventionalist subjectively assigns the score during the procedure, the subjective nature of this assignment could introduce variability. This can generate a range of reader assessments, which may differ based on their individual training and expertise. While patients with an mTICI score of 2c or 3 exhibit consistent scoring reliability, it's worth noting that there's a significant level of inter-reader variability for patients scoring mTICI 2b. This variability could potentially influence the interpretation of our model's predictive performance [HBR20]. The degree of recanalization experienced by patients scoring 2b on the modified Thrombolysis in Cerebral Infarction

(mTICI) scale can vary greatly, ranging from 50% to 89%. Consequently, the 2b score remains highly subjective. Since its inception in 2005, the TICI scoring system has undergone several modifications, largely in response to concerns regarding its variability and weak correlation with functional outcomes. Despite these modifications, including additional categories, mTICI 2b continues to cover a broad range of recanalization rates. Therefore, the pursuit to enhance the reliability and accuracy of this metric remains an ongoing endeavor. Future research could consider a reevaluation of these patients using the procedural imaging taken during Endovascular Thrombectomy (EVT). This could potentially allow for a more precise stratification of mTICI 2b patients, leading to a more nuanced quantification of recanalization rates. Routine clinical imaging protocols can introduce biases into the data that could hinder direct comparisons of imaging series inputs across different modalities. The current Magnetic Resonance (MR) stroke protocol generates an angiogram that covers only part of the brain, excluding lateral sections from the field-of-view. Similarly, the Computed Tomography (CT) protocol, optimized for patients presenting within 18 hours of symptom onset, captures perfusion imaging with an incomplete field-of-view along the superior/inferior axis. Due to existing treatment protocols, both series could not be registered using an image processing pipeline tailored for speed. While more sophisticated image registration techniques, including those that utilize Deep Learning (DL), might manage to perform partial registration, preliminary experiments suggest that such processes would be too time-consuming for practical application in real-world settings. Scanners at other institutions may have the capability to capture complete field-of-view angiography and perfusion imaging for both MR and CT studies. Cohorts with comprehensive coverage across all series could help to highlight the benefits of both standard and advanced imaging in predicting the success of Endovascular Thrombectomy (EVT). Lastly, this study serves as a proof-of-concept from a single institution and the architecture of the model involves numerous parameters. Despite the thoughtful split of training and evaluation cohorts to maximize evaluation capacity, and the utilization of two cohorts with different imaging modalities, the data is still drawn from

a single institutional dataset. Therefore, external validation is crucial to assess the wider applicability of these models across different hospitals and institutions.

4.6 Summary

We have presented a fully automatic, end-to-end method to predict treatment response to EVT. On a dataset of patients who received either MR or CT prior to treatment, we have demonstrated that the volume-based DL network can distinguish whether a patient will be successfully recanalized in one attempt or fewer, achieving peak accuracies of 88.80% using MR imaging and 82.33% using CT image series. This method outperformed previously published methods without requiring manual thrombus segmentation, illustrating the capability of DL algorithms to inform treatment planning for AIS patients. The future study includes a larger cohort for model training and a multicenter evaluation. With a larger training set, we can remove the preprocessing steps thus making the algorithm more applicable to real-world clinical settings.

CHAPTER 5

Large Vessel Occlusion Classification: A Masked Imaging Model Transformer Approach

5.1 Introduction

In Chapter 4, we discuss the CNN-Transformer hybrid model that is efficient for small datasets. However, highly customized hybrid transformer architecture is unsuitable for large-scale applications, notably because it diminishes the benefit of language-image unification. As the dataset grows, applying self-supervised learning such as Masked Imaging Model (MIM), also known as Masked Autoencoder (MAE) enables us to directly use pure transformer architecture that shows superior performance over the traditional CNN method when the dataset is large enough. The benefits of translational invariance and inductive bias introduced by convolution can be overcome by larger data and ultimately leads to better generalization ability.

5.2 Overview

Endovascular Thrombectomy (EVT) is the standard treatment for acute ischemic stroke (AIS) patients caused by large vessel occlusion (LVO) [GMV16]. The efficacy of EVT for patients experiencing AIS caused by LVO diminishes as time progresses [SGL16], which is optimal within 6 hours but still has some benefit within an extended treatment window from 6 hours to 24 hours. Many AIS patients are first assessed at regional centers that at most can

only manage intravenous thrombolysis, or on Mobile Stroke Units (MSU) that are equipped with CT machines that possess the ability to conduct both non-contrast CT (NCCT) and CT angiography (CTA) [CBS22]. Consequently, promptly detecting LVO and swiftly triaging patients to EVT-equipped comprehensive stroke centers (CSCs) is crucial. Although current guidelines recommend that primary stroke centers (PSCs) and other first-line facilities that provide initial emergency care include the administration of thrombolysis and the capability of performing emergency noninvasive intracranial vascular imaging such as CTA or MRA [PRA19, AKM19], many hospitals do not have CTA readily available [AAH19]. CTA is most widely used in many centers with very high sensitivity (83% to 97%) and specificity (87% to 99%) [LLR21, BJB20, FBH21]. On the MRI side, black-blood MRI is a unique modality that demonstrated high diagnostic accuracy and reliability with 100% sensitivity and specificity [AAE19]. Fluid Attenuated Inversion Recovery (FLAIR) also showed good sensitivity (98%) and specificity (86%) [BTL20]. However, the long-acquisition time and the lack of devices make MR a less favorable option for initial imaging screening. Even at well-equipped centers with CT machines, up to 20% of LVO may be missed at initial imaging evaluation when neuroradiologists are not available. When imaging devices are not presented for prescreening, many clinical scores can be used [NIM22]. The Los Angeles Motor Scale (LAMS) [NSS18], the Cincinnati Pre-hospital Stroke Scale (CPSS) [RHT18], the Rapid Arterial oCclusion Evaluation (RACE) scale [CCC17, LSD20], Glasgow Coma Scale (GCS) [HHJ21] are among the options. NIH Score Scale (NIHSS) is also modified and adopted for LVO prediction in multiple studies [PHR17, HHB16]. There are also several novel scales that were developed recently [TVC17, TSS19, VAF19]. However, the sensitivity and specificity are not comparable to imaging for triaging patients accurately and efficiently. Developing a Machine Learning (ML) algorithm capable of recognizing LVO patterns with the most common NCCT images could expedite and streamline the selection process for EVT candidates. By making this technology accessible to a broader array of primary stroke centers (PSCs), it can potentially expand the pool of patients receiving improved EVT treatment,

ultimately leading to better health outcomes. Czap et al. developed a Deep Learning (DL) model that uses CTAs obtained from 2 Mobile Stroke Units (MSUs) to detect LVO [CBS22]. Their proposed method DeepSymNet-v2 uses the information from the contralateral side in a siamese fashion using 3D CNN. Amukotuwa et al. developed an automated algorithm that relies on brain registration, region mapping, and ratio calculation with CTA [ASD19]. Another commercial software MethinksLVO used DL to predict LVO with NCCT [OCG20]. However, previous methods either mainly used CTA or have a balanced dataset, which is not aligned with real-world distribution and the performance in real-world settings needs to be evaluated. In this chapter, we propose to use the Masked Image Model (MIM) to pretrain a model on large NCCT data and use a 3D swin transformer (Swin-T) as the backbone to predict LVO.

5.3 Method

The LVO readings were determined by neuroradiologists through a visual assessment of CTA. The cohort used for this study was retrospectively collected from UCLA Ronald Reagan Medical Center from 2014-2021. AIS patients were included if they were diagnosed with a large-vessel occlusion (LVO) and had a CT exam upon admission under stroke protocol. non-LVO cases were included under the same stroke protocol but had non-LVO diagnosed, which may or may not have a stroke. Low image quality such as motion blur or spike is excluded. Unlike our previous work, we do not implement any preprocessing or image registration. Instead, we rely on extensive data augmentation for model training and thus loosen the restriction of the data and expand the generalization ability of the method.

5.3.1 CT Acquisition

Two CT scanners, a Lightspeed VCT (GE Health Care, Milwaukee, USA) and a SOMATOM Definition (Siemens, Forchheim, Germany), were used for NCCT and CTA imaging. After

administering 50 mL of contrast agent intravenously at 5 mL/second, a single-phase CT-angiography (CTA) was obtained (120 kV, 120 reference mAs, 0.3 second rotation time, 0.6 pitch, effective dose of about 3 mSv).

5.3.2 Implementation Details

We developed a 3D transformer architecture based on a 2D swin transformer (Swin-T) [LLC21]. To harvest the power of the transformer architecture, we adopted the SimpleMIM approach proposed in [XZC21] to a modified 3D version. SimpleMIM is inspired by Masked Autoencoder (MAE) but compatible with Swin-T given the vanilla MAE does not support the hierarchical architecture of Swin-T for the proposed masking strategy. Moreover, SimpleMIM demonstrated that a simple one-layer decoder is more than enough for effective self-supervised learning (SSL). Figure 5.1 illustrates the model's architecture and the MIM pretraining. It is worth noting that during MIM pertaining, the encoder is a pure 3D swin transformer. For downstream task LVO detection, we add a residual connection for the swin transformer layers and a convolution layer at the end to build a swin block. has shown this approach citeliang2021swinir to enhance the translational equivariance of the swin transformer. The residual connection allows better aggregation of multiscale features.

5.3.3 Evaluation metrics and Statistical Analysis

Categorical variables are shown in absolute numbers and percentages, and continuous variables as mean±SD or median with the interquartile range depending on the data type. Sensitivity, Specificity, Accuracy, Area Under the Precision-Recall Curve (PR-AUC), and Area under the receiver operating curves (ROC-AUC) are reported. The best cutoff for sensitivity and specificity is determined by Youden's J index. All statistical analysis was performed using R software 4.1.3 (https://www.r-project.org).

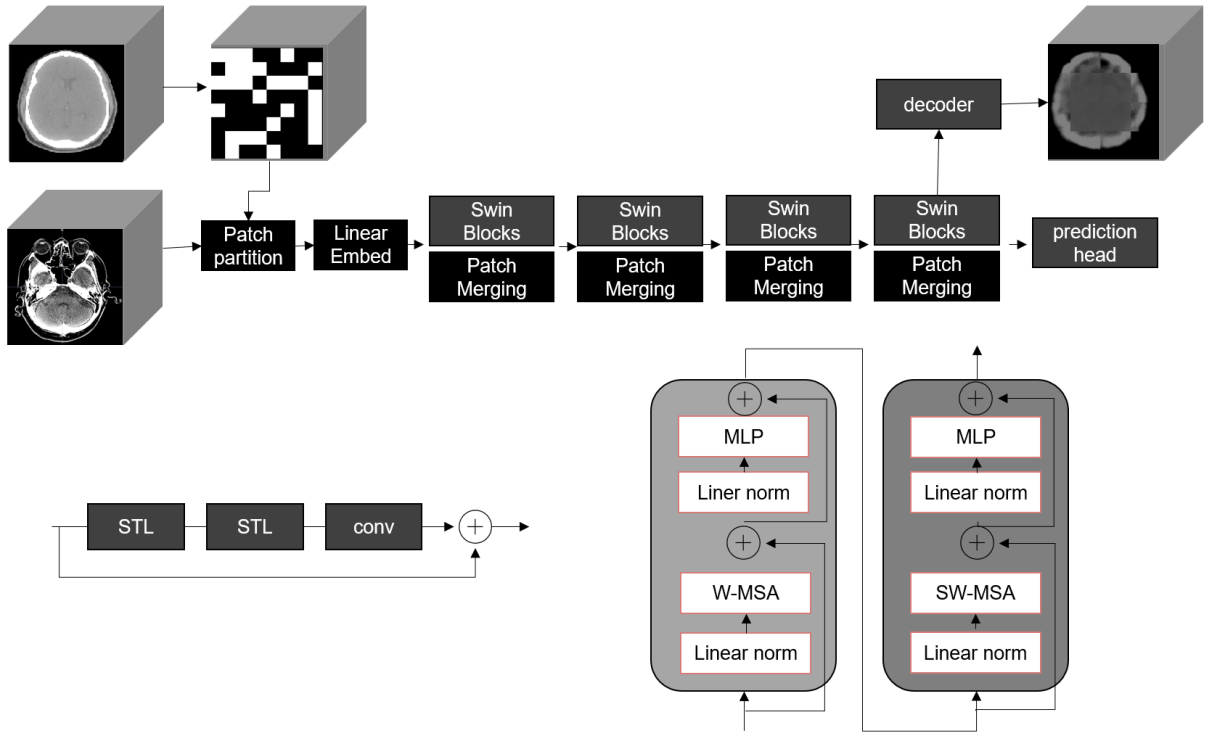


Figure 5.1: (a) Illustration of the proposed swin transformer LVO detection framework. STL stands for the swin transformer layer, which is detailed at the bottom right. Multiple STLs form a swin block. The input is 3D volume Non-contrast CT. The masking blocks are cubes for MIM.

5.4 Results

5.4.1 Materials

1,722 patients are included in this study based on the aforementioned inclusion criteria. Baseline characteristics are summarized in Table 5.1. 80% of the data are used for method development and 20% of the data construct a hold-out test set.

Dataset	(n = 1722)
Age (years)	70 (56-81)
Female	845 (49.07%)
LVO	327 (18.99%)
Stroke	400 (23.23%)

Table 5.1: Patient cohort basic demographics. Numbers are n (%) or median (interquartile ranges).

To illustrate the effectiveness of large sample sizes and self-supervised learning, we minimize the preprocessing steps. The only preprocessing is resampled to 128x128x128 to fit into the 3D Swin-T model. For the downstream LVO classification task, extensive data augmentation mechanics are used, including cropping, horizontal flipping, random noise, random elastic deformation, random affine transformation, and random anisotropy are used.

5.4.2 Impementation Details

The model is implemented using PyTorch 1.9 and trained on a DGX-1 with V100 GPUs. Mixed precision and accumulated steps are used. The learning rate is set to 0.0005 with a linear warmup strategy and cosine annealing learning rate scheduler. For MIM, the batch size is 1 and trained for 800 epochs; for the LVO task, the batch size is 26 and trained for 50 epochs. Focal loss is used to account for class imbalance problems in the data. The 3D Swin-T follows the following setup: the initial feature embedding size C is 48; The numbers

of swin transformer blocks at the four stages are 2, 2, 2, 2, with the corresponding numbers of attention head 3, 6, 12, 24. The patch size is 2x2x2 and the window size is 7x7x7. For MIM, the masking patch size is 16x16x16, and random patches strategy is adopted.

5.4.3 Model Performance

The performance metrics for comparing models are summarized in Table 5.2. When trained from scratch, 3D ResNet18 achieved the best performance with a ROC-AUC of 0.8698. 3D ViT is hard to converge and showed a low ROC-AUC of 0.7541. 3D Swin-T achieved a ROC-AUC of 0.8333 which is above 0.8 but still lower than 3D ResNet. However, when incorporating MIM, the performance of the ViT and Swin-T both increased, and the Swin-T achieved the best ROC-AUC of 0.8829. For ResNet18 3D, we use a Kinetic pre-trained model which is commonly used in other fields that applied 3D DL models but in the medical domain, these pre-trained weights showed no benefits. ROC comparison for 3D ResNet and Swin-T are shown in Figure 5.2

5.5 Conclusion

Our results demonstrate that our proposed transformer framework can accurately classify large vessel occlusion (LVO) in patients with suspected acute ischemic stroke (AIS) using only non-contrast CT (NCCT) images. The automated DL algorithm will accelerate the triage of EVT candidates and expand the population that would have been missed due to the limitation of CT or MR imaging under current guidelines. Accurate prediction of LVO using NCCT also remove the necessity of further advanced imaging exam, saving valuable time under emergent situation. It also improves the efficiency of prehospital, primary stroke center (PSC), and mobile stroke unit and sends the patients with LVO to the corresponding comprehensive stroke center (CSC). The proposed method aims to remove all the preprocessing steps that are commonly seen in other neuroimaging work but instead rely on more

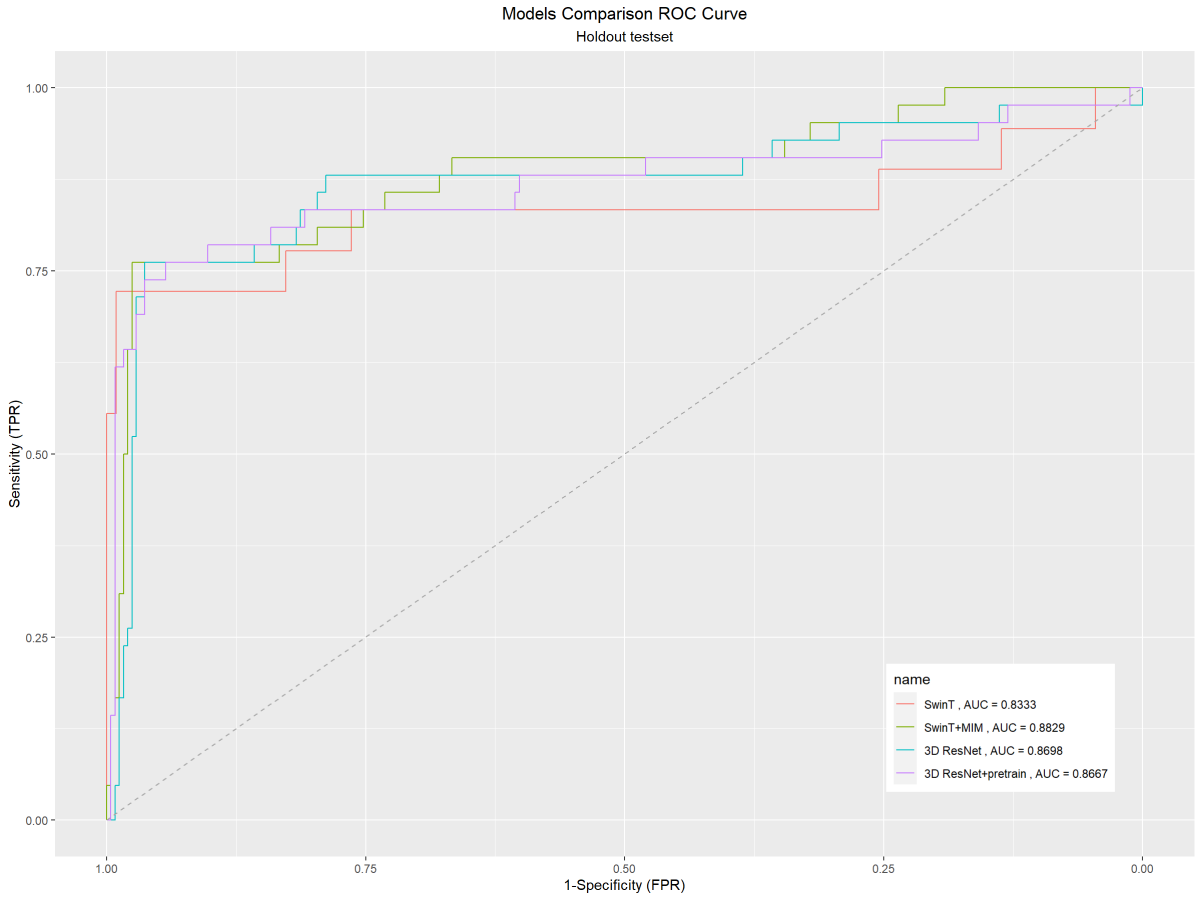


Figure 5.2: ROC curves for ResNet and Swin-T performance

Table 5.2: Quantitative evaluation of methods on the test set. The best results are in **bold**.

Models	Holdout test set			
	ROC-AUC	Accuracy	Sensitivity	Specificity
3D ResNet18	0.8698	0.9340	0.7619	0.9634
3D ViT	0.7541	0.8052	0.6675	0.7831
3D Swin-T	0.8333	0.9531	0.7222	0.9909
3D ReNet18+pretrain	0.8667	0.9167	0.7619	0.9431
3D ViT+pretrain	0.7904	0.8510	0.7034	0.8222
3D Swin-T+pretrain	**0.8829**	**0.9444**	**0.8421**	**0.9756**

data augmentation during training and the MIM pertaining, thus making the algorithm more inclusive and can be applied to a larger population as the preprocessing steps usually lead to some failed cases that need to be discarded. This is an exploratory study and therefore there are some limitations that should be addressed in future work. First, we only compared the resnet3D model and the Swin-T and their train-from-scratch and MIM performance. More comprehensive experiments should be included for ablation study and hyperparameter selection, such as masking strategy for MIM. This is a single-center study and a multicenter external validation study is essential to show the robustness of the proposed method.

5.6 Summary

In this chapter, we showed the transition from CNN/hybrid transformer work in previous studies to pure attention-based models and their superior performance when the masked imaging model pretraining is applied. This creates a new research path that may potentially

bridge the gap between Natural language processing and computer vision in the medical domain as the architectures are now unified therefore leading to better modality fusion. In the future, CLIP-alike architecture will exploit the information from the radiology report and the image simultaneously with no or limited annotations from expert radiologists, thus eventually leading to the development of a large foundation imaging model that can be easily fine-tuned or adapted to different downstream tasks in the medical domain.

CHAPTER 6

Transformer Volumetric Super-Resolution from CT Scans

6.1 Introduction

Chapter 4 presents a hybrid CNN-transformer architecture and 5 presents a pure transformer architecture for computer vision tasks. In this chapter, we further explore the swin transformer architecture and a unique application in super-resolution inspired by Masked Image Models (MIM). Unlike other MIM approaches that are designed for self-supervised learning for pretraining a model for downstream tasks, we directly apply the logic of mask reconstruction in MIM to super-resolution slice reconstruction from thick slice to thin slice in medical imaging. The work described in this chapter is partially in press as RPLHR-CT Dataset and Transformer Baseline for Volumetric Super-Resolution from CT Scans [YZK22].

6.2 Overview

Volumetric medical imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI), is an important tool in diagnostic radiology. Although high-resolution volumetric medical imaging provides more anatomical and functional details that benefit diagnosis [YHH20, XLG21, CHJ20], long acquisition time and high storage cost limit the wide application in clinical practice [HHM16]. As a result, it is routine to acquire anisotropic volumes in practice, which have high in-plane resolution and low through-plane resolution.

However, the disparity in resolution can lead to several challenges: (1) the inability to display sagittal or coronal views with adequate detail [KFW03]; (2) the insufficiency of spatial resolution to observe the details of lesions [YHH20] and; (3) the challenge to the robustness of 3D medical image processing algorithms [PLK21, IJK21]. A feasible solution is to use super-resolution (SR) algorithms [ZGD21] to upsample anisotropic volumes along the depth dimension, in order to restore high resolution (HR) from low resolution (LR). This approach is referred to as "volumetric SR."

CNN-based algorithms have achieved outstanding performance in SR for natural images [WCH20] and these techniques have been introduced for volumetric SR [PLL20, LZL20, GYX19, PZC21, CSC18, LLW21, LCX21, XSZ21, ZDP20, BLP18]. Though significant advances have been made, CNN-based algorithms remain limited by the inherent weaknesses of convolution operators. On the one hand, using the same convolution kernel to restore various regions may neglect content relevance. Liu et al. [LZL20] take this into consideration and propose a multi-stream architecture based on lung segmentation to recover different regions separately, but this is hard to be a one-size-fits-all solution. On the other hand, the non-local content similarity of images has been used as an effective prior in image restoration [ZZZ20]. Unfortunately, the local processing principle of the convolution operator makes the algorithms difficult to effectively model long-range dependence. Recently, transformer networks have shown good performance in several visual problems of natural image [DBK20, LLC21], including SR [CWG21, LCS21]. Self-attention mechanism is the key to the success of the transformer. Compared to CNN-based algorithms, the transformer can model long-range dependence in the input domain and perform dynamic weight aggregation of features to obtain input-specific feature representation enhancement [KNH21]. These results prompted us to explore a transformer-based SR method.

Another impediment to the application of volumetric SR methods is data. Most relevant studies use HR volume as ground truth and degrade it to construct paired pseudo-LR volumes with which to train and evaluate methods [PLL20, PZC21, CSC18, XSZ21, ZDP20]. For

instance, Peng et al. [PLL20] perform sparse sampling on the depth dimension of thin CT to obtain pseudo thick CT. Zhao et al. [ZDP20] simulate pseudo-LR MRI by applying an ideal low-pass filter to the isotropic T2-weighted MRI followed by an anti-ringing Fermi filter. However, the performance will be affected when tested on the real-LR volume [BLP18] because of the domain gap between pseudo- and real-LR volume. To avoid it, some studies collect real-paired LR-HR volumes [PLK21, LZL20, GYX19, LCX21, BLP18]. For example, Liu et al. [LZL20] collect 880 real pairs of chest CTs and construct a progressive upsampling model to reconstruct 1mm CT from 5mm CT. In the field of MRI, a large data set containing 1,611 real pairs of T1-weighted MRIs has been used to develop the proposed SCSRN method [LCX21]. However, a benchmark to objectively evaluate various volumetric SR methods is still lacking.

To address this deficiency, the first goal of this work is to curate a medium-sized dataset, named Real-Paired Low- and High-Resolution CT (RPLHR-CT), for volumetric SR. RPLHR-CT contains real-paired thin-CTs (slice thickness 1mm) and thick-CTs (slice thickness 5mm) of 250 patients. To the best of our knowledge, RPLHR-CT is the first benchmark for volumetric SR, which enables method comparison. The other goal of our work is to explore the potential of transformer architecture for volumetric SR. Specifically, we propose a novel Transformer Volumetric Super-Resolution Network (TVSRN). TVSRN is designed as an asymmetric encoder-decoder architecture with transformer layers, without any convolution operations. TVSRN is the first pure transformer used for CT volumetric SR. We re-implement and benchmark state-of-the-art CNN-based volumetric SR algorithms developed for CT and show that our TVSRN outperforms existing algorithms significantly. Additionally, TVSRN achieves a better trade-off between image quality, the number of parameters, and running time.

6.3 Dataset and Methodology

6.3.1 RPLHR-CT Dataset

Dataset Description. The RPLHR-CT dataset is composed of 250 paired chest CTs from patients. All data have been anonymized to ensure privacy. Philips machines were used to perform CT scans and the raw data were then reconstructed into thin CT (1mm) and thick CT (5mm) images. Thus, recovering thin CT (HR volume) from thick CT (LR volume) for this dataset is a volumetric SR task with an upsampling factor of 5 in the depth dimension. The CT scans are saved in NIFTI (.nii) format with volume sizes of $L \times 512 \times 512$, where 512×512 is the size of CT slices, and L is the number of CT slices, ranging from 191 to 396 for thin CT and 39 to 80 for thick CT. The thin CT and the corresponding thick CT have the same in-plane resolution, ranging in $[0.604, 0.795]$, and are aligned according to spatial location.

Dataset split and Evaluation Metric. We randomly split the RPLHR-CT dataset into 100 train, 50 validation and 100 test CT pairs. For evaluation, we quantitatively assess the performance of all methods in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [WBS04]. Significance is tested by one-sided Wilcoxon signed-rank test.

Dataset Analysis. To analyze the difference between the thin CT and thick CT, we group slices in thin CT and thick CT into three categories of slice-pairs according to their spatial relationship, as shown on the left side of Fig. 6.1. We use PSNR and SSIM to access the changes in the similarity of three slice pairs in train, validation, and test CT pairs. As shown on the right side of Fig. 6.1, the results indicate that the similarity of slice-pairs at the same spatial location in thin CT and thick CT, namely **Match**, is the highest, while the similarity decreases significantly as the spatial distance becomes larger.

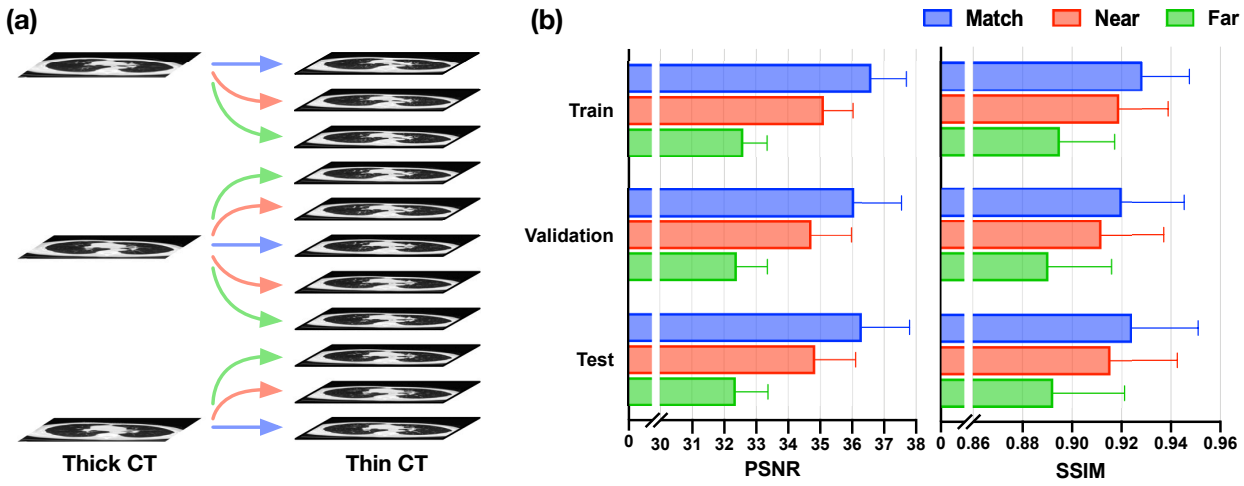


Figure 6.1: (a) Three categories of slice-pairs according to their spatial relationship in thin CT and thick CT. Match: same position, shown in blue; Near: 1mm apart, shown in red; Far: 2mm apart, shown in green. (b) The degree of similarity between the three slice-pairs on the three datasets. (Color figure online)

6.3.2 Network Architecture

Inspired by MAE[HCX21], we treat volumetric SR as a task to recover the masked regions from the visible regions, where the visible regions refer to the slices in the LR volume and the masked regions refer to the slices in the corresponding HR volume. As illustrated in Fig. 6.2, we also design our TVSRN with an asymmetric encoder-decoder architecture, but with several targeted modifications. First, in TVSRN, the encoder and the decoder are equally important, and to better model the relationship between the visible regions and the masked regions, the decoder uses a larger amount of parameters than the encoder. Second, instead of the standard transformer layer[DBK20], we use the swin transformer layer (STL)[LLC21], which is less computationally intensive and more suitable for high-resolution image, as the basic component of TVSRN. Third, we propose Through-plane Attention Blocks to exploit the spatial positional relationship of volumetric data to achieve better performance.

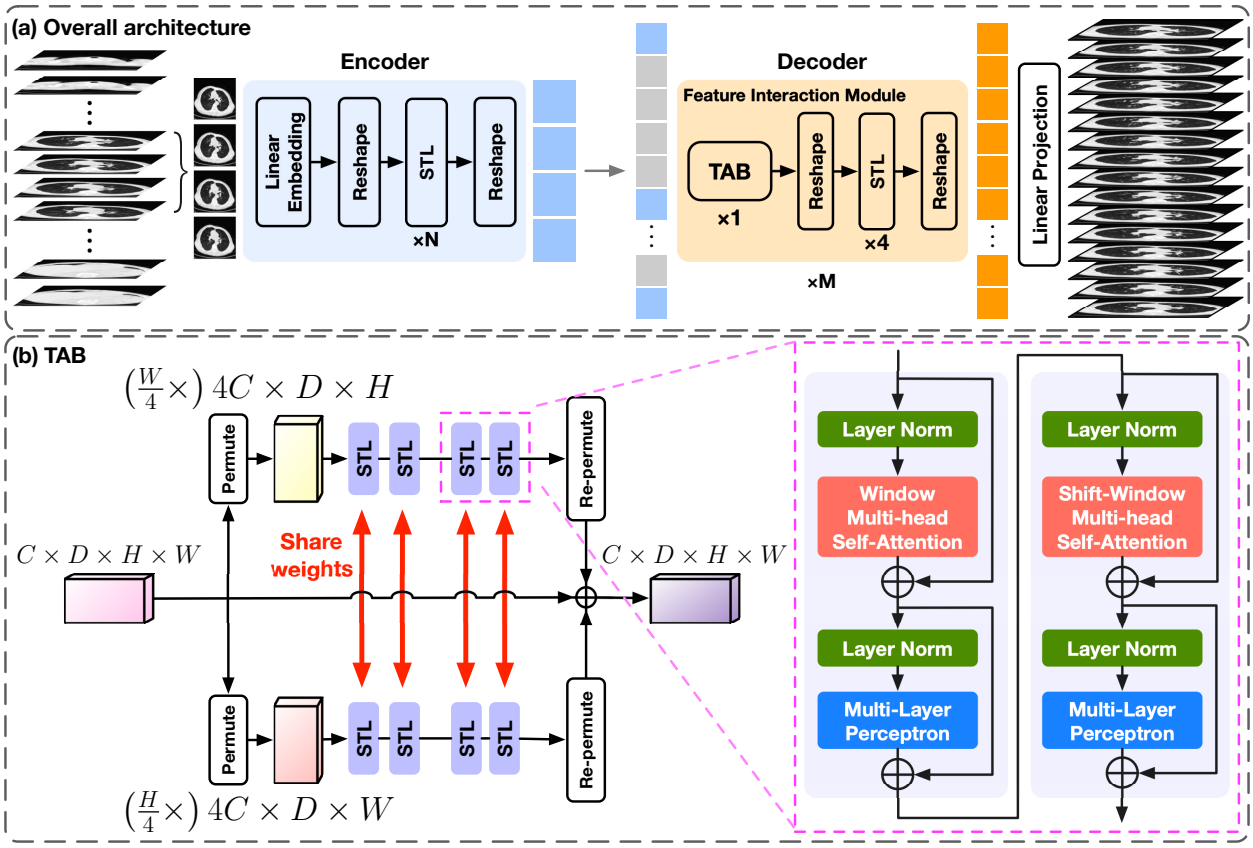


Figure 6.2: (a) Illustration of the proposed Transformer Volumetric Super-Resolution Network architecture. (b) Details of TAB. The purple dashed box represents two consecutive swin transformer layers. The batch dimension is indicated in parentheses.

Encoder is used to map the LR volume to a latent representation. The consecutive slices from LR volumes are denoted as the input $X_e^{in} \in \mathbb{R}^{1\times D\times H\times W}$ of *encoder*, where D, H and W are the depth, height, and width, and the channel is 1. X_e^{in} is firstly fed into the *Linear Embedding*, whose number of feature channel is C, to extract shallow features and output $F_s \in \mathbb{R}^{C\times D\times H\times W}$. Then, F_s is reshaped to $F_0 \in \mathbb{R}^{CD\times H\times W}$. We stack N STLs to extract deep features from F_0 as:

$$F_i = H_i^{STL}(F_{i-1}), \quad i = 1, 2, ..., N \tag{6.1}$$

where $H_i^{STL}(\cdot)$ denotes the i-th STL. Finally, F_N is reshaped to 3D output $X_e^{out} \in \mathbb{R}^{C\times D\times H\times W}$.

Decoder is used to recover the HR volume from the latent representation. As shown in Fig. 6.2(a), mask tokens are introduced after the *encoder*, and the full set of X_e^{out} and mask tokens is input to the *decoder* as $X_d^{in} \in \mathbb{R}^{C\times D'\times H\times W}$, where D' is the depth of ground truth. The mask tokens are a learned vector that indicates the missing slices in the LR volumes compared to the HR counterpart. *Decoder* stack M Feature Interaction Modules (FIMs), which consists of one Through-plane Attention Block (TAB), four STLs, and two reshape operations. The reshape operations are used to reshape the input feature map into the size expected by the next block. The output of the *decoder* is X_d^{out} with the same size as X_d^{in}. Note that the design of asymmetric *decoder* can easily be adapted to other upsampling rates by changing the number of mask tokens.

The details of TAB are illustrated in Fig. 6.2(b). TAB is the first block in each FIM. There are two parallel branches in TAB that perform self-attention on the input from coronal and sagittal views, respectively. In both views, the depth dimension will become an axis of the STL's window, so the relative position relationship between slices will be incorporated into the calculation. The parameter weights of the corresponding STL on the two parallel

branches are shared. Given the input feature z_{in} of TAB, the output is computed as:

$$
\begin{aligned}
z_0^{sag} = P^{sag}(z_{in}),\ z_0^{cor} = P^{cor}(z_{in}) \\
z_j^{sag} = H_j^{STL}(z_{j-1}^{sag}),\ z_j^{cor} = H_j^{STL}(z_{j-1}^{cor}),\ \ j = 1, 2, 3, 4 \\
z_{out} = z_{in} + P_{re}^{sag}(z_4^{sag}) + P_{re}^{cor}(z_4^{cor})
\end{aligned} \tag{6.2}
$$

where $P^{sag}(\cdot)$ and $P^{cor}(\cdot)$ are permutation operations that transform the input to sagittal and coronal view, respectively. $P_{re}^{sag}(\cdot)$ and $P_{re}^{cor}(\cdot)$ denote re-permutation operations that reshape the input back to its original size. In addition, TAB contains residual connections, which allow the aggregation of different levels of features.

Reconstruction Target. The X_d^{out} is fed into the *Linear Projection* to obtain the pixel-wise prediction $\hat{Y} \in \mathbb{R}^{D' \times H \times W}$. The L_1 pixel loss is formulated as:

$$
L_{pixel} = \frac{1}{D' \times H \times W} \sum_{k,i,j} |\hat{Y}_{k,i,j} - Y_{k,i,j}| \tag{6.3}
$$

where Y is the ground truth HR volume.

Architecture Hyper-parameters. For each STL, the patch size is 1×1 and the window sizes of the x-axis, y-axis, and z-axis are set to 8, 8, and 4. For *Linear Embedding*, the channel number C is 8. The number of STLs in the encoder and FIMs in the decoder is set to $N = 4$ and $M = 1$, respectively.

6.4 Experiments and Results

Implementation Details. We normalize the intensity of the CT images from $[-1024, 2048]$ to $[0, 1]$. During training, $4 \times 256 \times 256$ cubes from thick CTs are used as input and the corresponding $16 \times 256 \times 256$ cubes from thin CTs are used as ground truth, in where $16 = (4 - 1) \times 5 + 1$. During inference, we feed cubes from thick CTs to the model in a sliding window manner, in which the overlap of depth dimension is 1 and the rest is 0. If the depth of untested cubes is less than 4, we feed the last 4 slices into the model. For

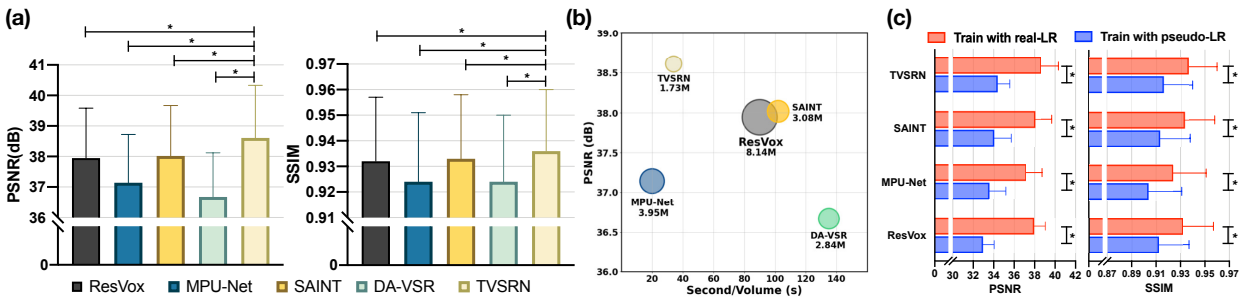


Figure 6.3: (a) Quantitative comparisons of our TVSRN and other state-of-the-art methods. $*$ indicates $p < 0.001$. (b) PSNR vs. processing time of each volume with the number of parameters shown in circle size. (c) quantitative results of pseudo images experiment.

multiple predictions on the same coordinate, we take the average as the final value. TVSRN is trained with Adam Optimizer. The learning rate is 0.0001 and the batch size is 1. For the comparison methods, we follow descriptions provided in the original papers to re-implement the models, as none have public code available. Settings not detailed in the original paper will remain consistent with our work. Data augmentation includes random cropping and horizontal flipping. The framework is implemented in PyTorch and trained on NVIDIA A6000 GPUs.

6.4.1 Results and Analysis

Fig. 6.3(a) summarizes the quantitative comparisons of our method and other state-of-the-art CT volumetric SR methods: ResVox [GYX19], MPU-Net [LZL20], SAINT [PLL20] and DA-VSR [PZC21]. For ResVox, the noise reduction part is removed. For MPU-Net, we do not use the multi-stream architecture due to the lack of available lung masks. TVSRN achieves PSNR of 38.609 ± 1.721 and SSIM of 0.936 ± 0.024 outperforms others significantly ($p < 0.001$). Moreover, as shown in Fig. 6.3(b), compared to other methods, TVSRN achieves a better trade-off in terms of the PSNR (optimal), the number of parameters (optimal), and the running time (suboptimal). We also perform the comparison on an external test set,

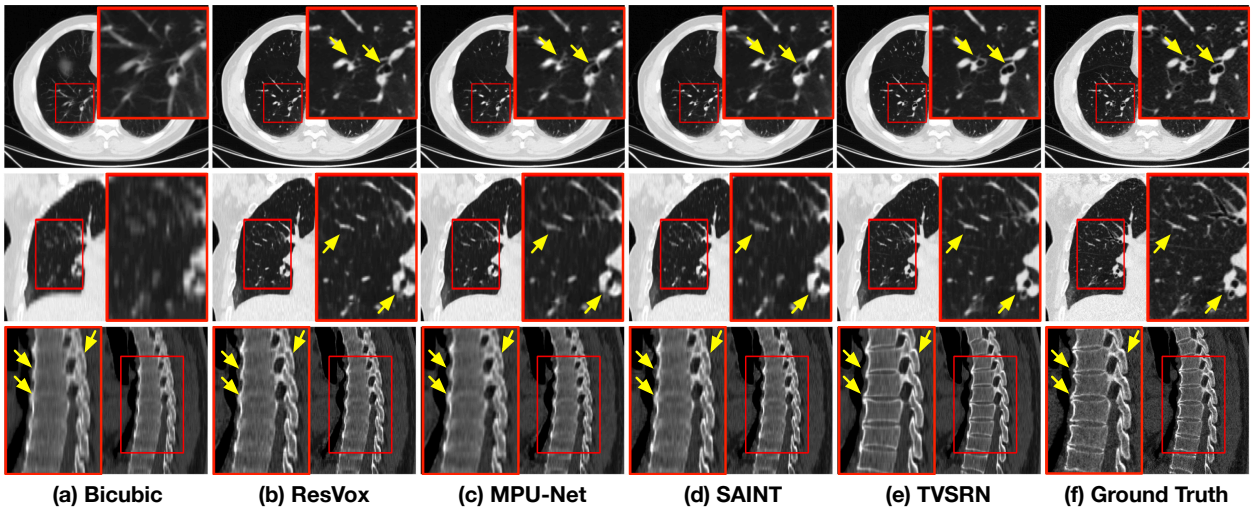


Figure 6.4: Visual comparisons of different methods against TVSRN. The first and second rows show the axial view and coronal view respectively, displayed as lung window. The third row is sagittal view, displayed as bone window. Yellow arrows point to areas of marked difference.

where TVSRN also achieved the best performance. Detailed numerical results on the internal test set and external test set are presented in the supplementary material. In addition, a sample-by-sample performance scatterplot is given in the supplementary material.

We visualize the axial, coronal, and sagittal views of HR CT volume obtained by different methods. It is clear in Fig. 6.4 that TVSRN has the richest details and the least amount of structural artifacts remaining in different views.

6.4.2 Domain Gap Analysis

We conduct a pseudo images experiment to illustrate the effect of the domain gap. Specifically, we degrade the training data to obtain pseudo-LR volumes and use these data to train several different methods. All settings are the same as those in the previous section, except for the training data. For testing, real-LR volumes in the internal test set are used as input

to calculate the PSNR and SSIM. As shown in Fig. 6.3(c), the results show that both PSNR and SSIM of various methods are significantly decreased to varying degrees ($p < 0.001$). Please refer to the supplemental material for more details on degradation.

6.4.3 Ablation Study

The ablation study is used to verify the contribution of each component in TVSRN on performance. The full TVSRN is compared to:

- $\text{TVSRN}_{ViT}^{Encoder}$. A standard transformer-based method based on [DBK20]. We map each patch of size $1 \times 16 \times 16$ to a token with a length of 512 and set the number of transformer layers to eight. Instead of asymmetric *decoder*, it uses subpixel conversion [SCH16] to perform upsampling.

- $\text{TVSRN}^{Encoder}$. Only the *encoder* of TVSRN was used. N is increasd to eight and C is increased to 32. The upsampling method is subpixel convert.

- $\text{TVSRN}^{w/o\ TAB}$. TAB is not used in TVSRN, that is, the relative position relationship among slices is ignored in the network.

Model performance is summarized in Table. 6.1. Notable observations include: 1) among all designs, $\text{TVSRN}_{ViT}^{Encoder}$ has the most parameters but the worst performance, which indicates that it is not feasible to simply apply the transformer to the volumetric SR; 2) replacing the standard transformer layer with STL can greatly reduce the number of parameters and improve the performance by a large margin (up to 2.827dB); 3) asymmetric decoder can improve performance slightly without changing the number of parameters; 4) improvements can be seen from $\text{TVSRN}^{w/o\ TAB}$ to TVSRN, indicating the effectiveness of modeling the relative position relationship among slices. Sample-by-sample performance scatterplots in supplemental material are used to further illustrate the effectiveness of individual components.

Table 6.1: Results of ablation study for TVSRN in terms of PSNR and SSIM. The best results are **bolded**, and the second best results are underlined. * denotes statistically significant ($p < 0.001$) against the above method with a one-sided Wilcoxon signed-rank test.

Designs	Param	PSNR(↑)	SSIM(↑)
$\text{TVSRN}_{ViT}^{Encoder}$	17.15M	35.537 ± 1.353	0.918 ± 0.026
$\text{TVSRN}^{Encoder}$	1.58M	$38.364 \pm 1.675^*$	$0.934 \pm 0.024^*$
$\text{TVSRN}^{w/o\ TAB}$	1.56M	$\underline{38.497 \pm 1.700^*}$	$\underline{0.935 \pm 0.024^*}$
TVSRN	1.73M	$\mathbf{38.609 \pm 1.721^*}$	$\mathbf{0.936 \pm 0.024^*}$

6.5 Conclusion

A persistent problem with volumetric SR is the lack of real-paired data for training and evaluation, which makes it challenging to generalize algorithms to real-world datasets for practical applications. In this paper, we presented the RPLHR-CT Dataset, which is the first open real-paired dataset for volumetric SR, and provided baseline results by re-implementing four state-of-the-art SR methods. We also proposed a convolution-free transformer-based network, which significantly outperformed existing CNN-based methods and has the least number of parameters and the second shortest running time. In the future, we will enlarge the RPLHR-CT Dataset and investigate new volumetric SR training strategies, such as semi-supervised learning or using unpaired real data.

6.6 Appendix

6.7 Internal Test Set

Numerical results are summarized in Table. 6.2. Sample-by-sample performance (PSNR and SSIM) scatterplots for the first 15 cases are shown in Fig. 6.5.

Table 6.2: Quantitative evaluation of methods on the internal test set. The best results are in **bold**. 95% confidence intervals are in square brackets. * denotes the statistically significant difference ($p < 0.001$ in one-sided Wilcoxon signed-rank test) between the current method and TVSRN.

Models	Internal test set	
	PSNR(↑)	SSIM(↑)
Bicubic	$33.508 \pm 1.083^*$ $[31.628, 35.225]$	$0.902 \pm 0.028^*$ $[0.844, 0.947]$
ResVox	$37.946 \pm 1.637^*$ $[35.527, 40.306]$	$0.932 \pm 0.025^*$ $[0.888, 0.968]$
MPU-Net	$37.140 \pm 1.583^*$ $[34.854, 39.457]$	$0.924 \pm 0.027^*$ $[0.874, 0.963]$
SAINT	$38.019 \pm 1.651^*$ $[35.565, 40.515]$	$0.933 \pm 0.025^*$ $[0.890, 0.969]$
DA-VSR	$36.672 \pm 1.450^*$ $[34.469, 38.760]$	$0.924 \pm 0.026^*$ $[0.876, 0.961]$
TVSRN	$\mathbf{38.609 \pm 1.721}$ $\mathbf{[36.047, 41.282]}$	$\mathbf{0.936 \pm 0.024}$ $\mathbf{[0.895, 0.970]}$

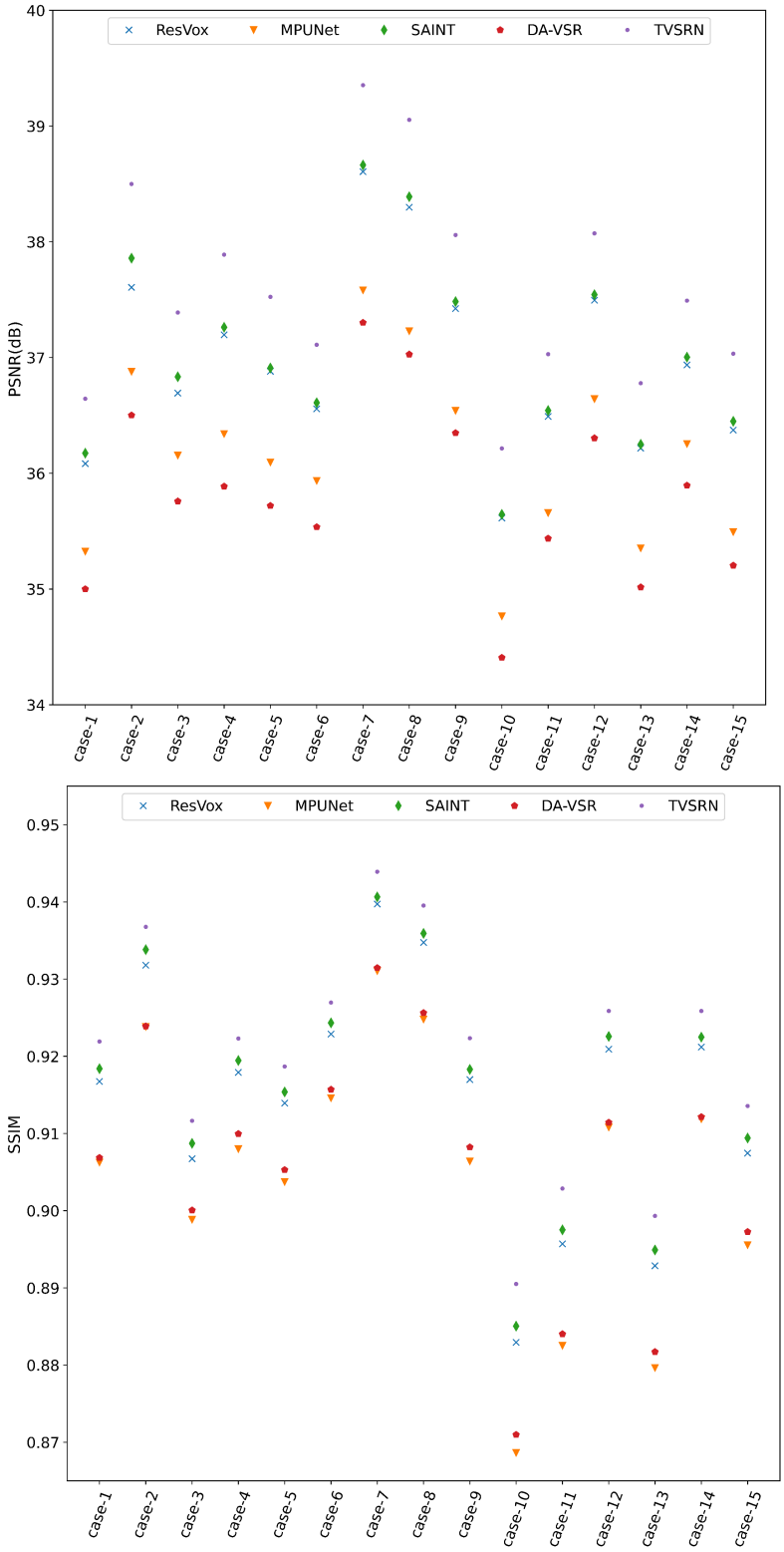


Figure 6.5: Sample-by-sample performance scatterplot on the internal test set.

6.8 External Test Set

We collected 18 chest CT pairs from another center as an external test set, which were obtained using Toshiba scanners and reconstructed to thin slice CT (1mm) and thick slice CT (5mm) volumes. All data have been anonymized. Numerical results are summarized in Table. 6.3. Sample-by-sample performance (PSNR and SSIM) scatterplot for all cases in the external test set are shown in Fig. 6.6.

6.9 Summary

In this chapter, we create a novel way to use Swin Transformer for the Super Resolution task. Given the limitation of real-paired data, we can only find lung CT data for this study. In stroke imaging, it is in fact very important to use volumetric super-resolution to enhance the image since for stroke protocol it is usually 5mm thick slice images. Many AI algorithms are developed to work on isotropic images and the volumetric super-resolution algorithm expands the utility of various AI algorithms. In future studies, we plan to use unpaired data to fine-tune this model to adapt to different organs and modalities.

Table 6.3: Quantitative evaluation of methods on the external test set. The best results are in **bold**. 95% confidence intervals are in square brackets. * denotes the statistically significant difference ($p < 0.001$ in one-sided Wilcoxon signed-rank test) between the current method and TVSRN.

Models	External test set	
	PSNR(↑)	SSIM(↑)
Bicubic	$31.592 \pm 1.133^*$ $[30.288, 33.551]$	$0.902 \pm 0.035^*$ $[0.843, 0.945]$
ResVox	$34.445 \pm 1.683^*$ $[32.339, 37.290]$	$0.927 \pm 0.032^*$ $[0.874, 0.965]$
MPU-Net	$34.012 \pm 1.454^*$ $[32.175, 36.389]$	$0.918 \pm 0.033^*$ $[0.863, 0.958]$
SAINT	$34.719 \pm 1.634^*$ $[32.716, 37.481]$	$0.929 \pm 0.031^*$ $[0.878, 0.967]$
DA-VSR	$33.296 \pm 1.458^*$ $[31.458, 35.744]$	$0.911 \pm 0.032^*$ $[0.860, 0.950]$
TVSRN	$\mathbf{35.720 \pm 1.708}$ $\mathbf{[33.644, 38.399]}$	$\mathbf{0.932 \pm 0.030}$ $\mathbf{[0.880, 0.968]}$

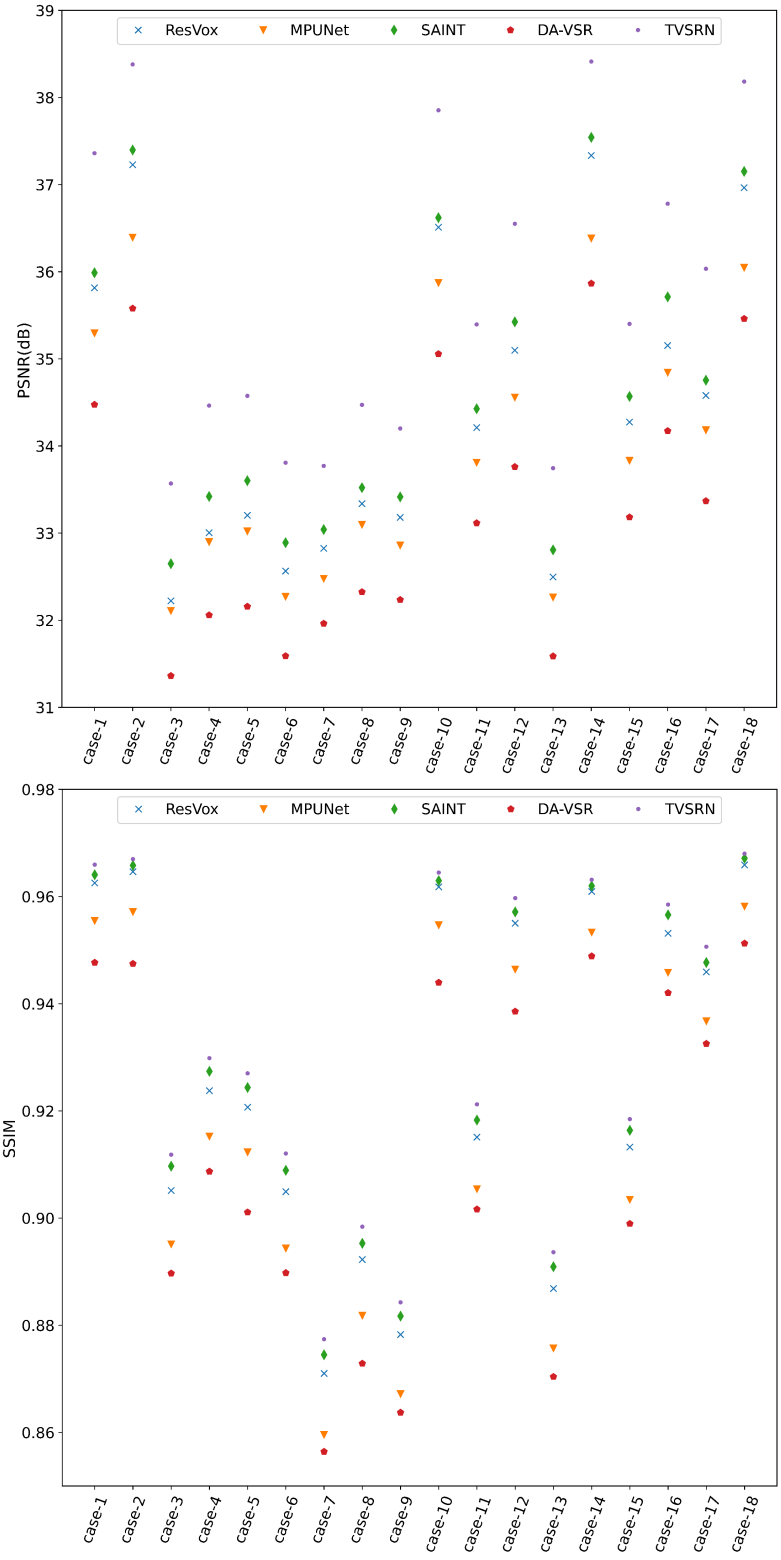


Figure 6.6: Sample-by-sample on the external test set.

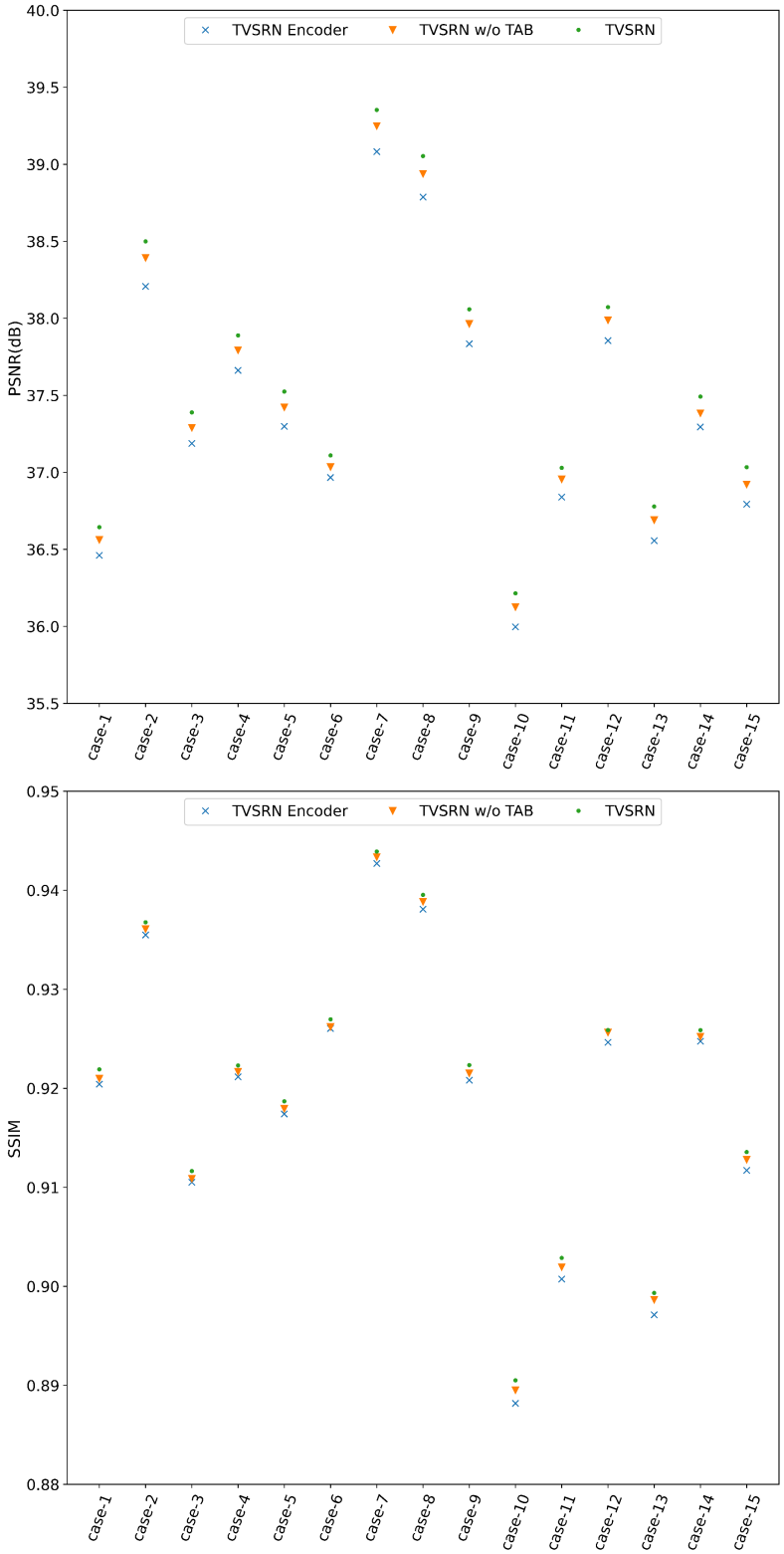


Figure 6.7: Sample-by-sample performance scatterplot on the internal test set of ablation study.

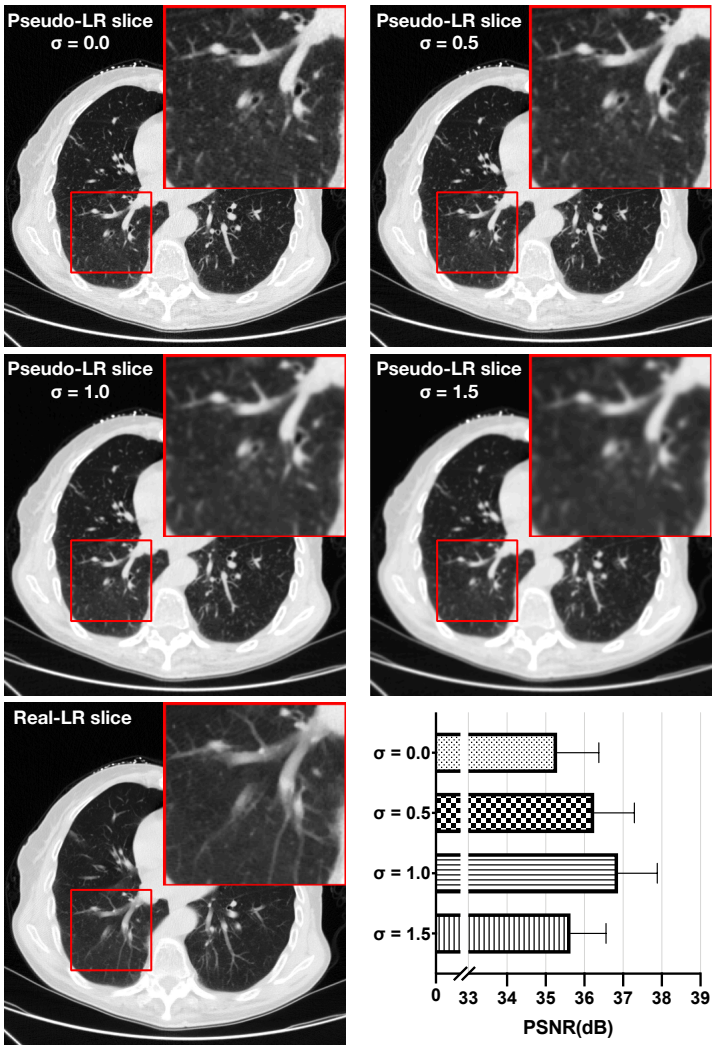


Figure 6.8: Comparison of different degradation strategies. First use bicubic interpolation to downsample the thin-CT to the same number of slices as the thick-CT, then perform Gaussian filtering. Four σ were set for the Gaussian filter, 0, 0.5, 1.0, and 1.5. When the $\sigma = 0$, it means that Gaussian filtering is not performed. Using peak signal-to-noise ratio (PSNR) to compare the similarity between pseudo-LR volumes and real-LR volumes obtained by four different degradation strategies, the results are shown in the lower right corner. When $\sigma = 1.0$, the pseudo-LR volume has the highest PSNR with the real-LR volume, but it still has a visible difference in appearance.

CHAPTER 7

Conclusion and Future Work

This Chapter summarizes the main contributions of this dissertation and discusses potential future directions to investigate.

7.1 Summary of Contributions

This dissertation presents models that utilize self-supervised learning pretraining and attention mechanisms to tackle the challenges of limited medical images for Deep Learning (DL) models to converge. The dissertation also expands the utility of DL model applications that are limited by thick slice images using transformer-based super-resolution preprocessing.

1. In Chapter 3, an intra-domain task-specific self-supervised learning approach and attention-based 2D and 3D CNN models to classify time since stroke using diffusion-weighted MRI. The proposed self-supervised learning approach significantly improves the TSS classification using pretreatment MRI, demonstrating that using only the intra-domain medical imaging for the task without imageNet or other unrelated medical imaging pretraining, the model can still reach high performance.

2. In Chapter 4, a CNN-transformer hybrid model to predict EVT outcomes using both CT and MRI was presented. we first developed radiomics-based Machine Learning (ML) models to examine the feasibility of predicting mTICI using pre-treatment MRI. Second, we develop a CNN-transformer hybrid model to predict mTICI using non-contrast CT and CT angiography. finally, we further add contrastive self-supervised

learning pretraining to the model and evaluate the performance on both CT and MRI. This chapter demonstrated a process to use DL to explore the relationship between medical images and EVT outcomes.

3. In Chapter 5, A pure vision transformer model is developed to predict large vessel occlusion using only non-contrast CT images. The proposed method used an improved 3D swin transformer model to accelerate the training. a 3D modified Masked imaging self-supervised learning approach to pretrain the model that demonstrates its effectiveness. The model achieved state-of-the-art performance using a real-world distributed dataset.

4. In Chapter 6, A transformer-based super-resolution model to synthesize 1mm slice thickness from 5mm slice thickness has been developed. We also released 800 real-paired 1mm and 5mm CT image volumes. We first demonstrate the importance of using real-paired data for super-resolution models compared to using real 1mm and downsampled 5mm images as training data. We then developed a swin transformer super-resolution model to achieve the state-of-the-art performance.

7.2 Future Work

Machine Learning (ML) and Deep learning (DL) possesses the capacity to significantly advance stroke management by facilitating faster, more accurate, and highly efficient diagnosis and treatment of Acute Ischemic Stroke (AIS). As the development of DL algorithms progress, medical imaging is paving the way for more tailored, patient-specific diagnostic and therapeutic approaches. These DL algorithms excel at rapidly and accurately analyzing extensive datasets, enabling them to detect patterns in medical imaging that are unidentifiable to the naked eye. The ongoing advancements in the wider domains of ML and DL offer a wealth of opportunities for technological innovation, harnessing methods that are optimally suited for handling medical data and addressing clinical challenges. Particularly,

this dissertation demonstrated the performance and the expansion of the capability of DL through the technical advancement in attention mechanism and self-supervised learning over recent years. By shifting from convolutional neural network (CNN) to attention-based CNN to transformers architecture, the gap between natural language processing (NLP) and computer vision (CV) has finally been closed. The large language model (LLM) has already shown its power in NLP and the large imaging model (LIM) is on the horizon. By linking LLM and LIM, the true potential of DL will be unleashed with large enough multimodal datasets.

On the hand, unlike natural imaging and natural language, the medical domain has its unique challenges. Many clinical problems are intrinsically non-deterministic, making it hard to model by simply feeding more data and using larger models. Medical imaging modalities, such as CT, MRI, and Ultrasound do not exist in the natural world and unlike natural images, are very different across modalities. Specially tailored methods need to be designed to work for different modalities and different diseases. In the foreseeable future, there still remain a lot of challenges to create a reliable single large foundation model that works for all or most of the situations in clinical settings.

While the long-term goals discussed above seem very challenging, the short-term goals still are to improve sample size and expand external validation, reduce specially tailored blocks/modules/networks, and expand the utilization of DL for different clinical problems. Meanwhile, effectively integrating robust algorithms into the current clinical workflow that improve real-world diagnostics accuracy and efficiency would attract more investment into the research in DL applications for the medical domain.

This dissertation focus on the development of DL algorithms with a core consideration of incorporating the algorithm to solve a real clinical problem and improve the current workflow. The utilization of attention mechanisms and self-supervised learning to make the DL algorithms work robustly with the major limitation of data size, which is common in the medical imaging domain, particularly for rare diseases or fewer studies topics. The

dissertation also showed a trend to not modify the network itself too much. This is crucial given the DL community is building larger and more efficient foundation models that can be fine-tuned for downstream tasks. These large models are shown to be much more robust and generalizable than small models that are highly customized and trained on small datasets. Although it is still necessary to add some modules and loss functions that may improve the downstream tasks, it is important to have a backbone that can be directly transferred from other well-trained general models. In the next step, although building a foundation model for the entire medical field is extremely challenging, a foundation model that only focuses on MRI or CT neuroimaging that use a combination of in-house and the public dataset is warranted and will provide valuable experimental results for the community to continue the development of large foundation models.

In addition, this dissertation mainly focuses on the development and validation of deep-learning models for medical imaging. The reasoning behind the shift from CNN to transformers is not only the superior performance when the training data is large but more importantly it will create a unified architecture for CV and NLP, making modality fusion more standardized. To this end, in the next step, we should exploit the rich text information from radiology report, patients report, and other text-related clinical information to mix with imaging features to achieve a more reliable diagnosis and treatment prediction.

REFERENCES

[AAE19] Anas S Al-Smadi, Ramez N Abdalla, Ali H Elmokadem, Ali Shaibani, Michael C Hurley, Matthew B Potts, Babak S Jahromi, Timothy J Carroll, and Sameer A Ansari. "Diagnostic accuracy of high-resolution black-blood MRI in the evaluation of intracranial large-vessel arterial occlusions." *American Journal of Neuroradiology*, **40**(6):954–959, 2019.

[AAH19] Sami Al Kasab, Eyad Almallouhi, Jillian Harvey, Nancy Turner, Ellen Debenham, Juanita Caudill, Christine A Holmstedt, and Jeffrey A Switzer. "Door in door out and transportation times in 2 telestroke networks." *Neurology: Clinical Practice*, **9**(1):41–47, 2019.

[Ahm13] Niaz Ahmed. "Results of Intravenous Thrombolysis Within 4.5 to 6 Hours and Updated Results Within 3 to 4.5 Hours of Onset of Acute Ischemic Stroke Recorded in the Safe Implementation of Treatment in Stroke International Stroke Thrombolysis Register (SITS-ISTR)." *JAMA Neurology*, **70**:837, 7 2013.

[AKM19] MA Almekhlafi, WG Kunz, BK Menon, RA McTaggart, MV Jayaraman, BW Baxter, D Heck, D Frei, CP Derdeyn, T Takagi, et al. "Imaging of patients with suspected large-vessel occlusion at primary stroke centers: available modalities and a suggested approach." *American Journal of Neuroradiology*, **40**(3):396–400, 2019.

[AMK18] Gregory W. Albers, Michael P. Marks, Stephanie Kemp, Soren Christensen, Jenny P. Tsai, Santiago Ortega-Gutierrez, Ryan A. McTaggart, Michel T. Torbey, May Kim-Tenser, Thabele Leslie-Mazwi, Amrou Sarraj, Scott E. Kasner, Sameer A. Ansari, Sharon D. Yeatts, Scott Hamilton, Michael Mlynash, Jeremy J. Heit, Greg Zaharchuk, Sun Kim, Janice Carrozzella, Yuko Y. Palesch, Andrew M. Demchuk, Roland Bammer, Philip W. Lavori, Joseph P. Broderick, and Maarten G. Lansberg. "Thrombectomy for Stroke at 6 to 16 Hours with Selection by Perfusion Imaging." *New England Journal of Medicine*, **378**, 2 2018.

[ASD19] Shalini A Amukotuwa, Matus Straka, Seena Dehkharghani, and Roland Bammer. "Fast automatic detection of large vessel occlusions on CT angiography." *Stroke*, **50**(12):3431–3438, 2019.

[ATS09] Brian B Avants, Nick Tustison, Gang Song, et al. "Advanced normalization tools (ANTS)." *Insight j*, **2**(365):1–35, 2009.

[AVC21] A.M. Alexandre, I. Valente, A. Consoli, M. Piano, L. Renieri, J.D. Gabrieli, R. Russo, A.A. Caragliano, M. Ruggiero, A. Saletti, G.A. Lazzarotti, M. Pileggi, N. Limbucci, M. Cosottini, A. Cervo, F. Viaro, S.L. Vinci, C. Commodaro, F. Pilato, and A. Pedicelli. "Posterior Circulation Endovascular Thrombectomy for

Large-Vessel Occlusion: Predictors of Favorable Clinical Outcome and Analysis of First-Pass Effect." *American Journal of Neuroradiology*, **42**:896–903, 5 2021.

[BAD96] Michael Brant-Zawadzki, Dennis Atkinson, Mark Detrick, William G Bradley, and Gerald Scidmore. "Fluid-attenuated inversion recovery (FLAIR) for assessment of cerebral infarction: initial clinical experience in 50 patients." *Stroke*, **27**(7):1187–1191, 1996.

[Ban11] O Y Bang et al. "Collateral flow predicts response to endovascular therapy for acute ischemic stroke." *Stroke*, **42**:693–699, 3 2011.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473*, 2014.

[BDP21] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. "BEiT: BERT Pre-Training of Image Transformers." 6 2021.

[BDS16] Vinit Baliyan, Chandan J Das, Raju Sharma, and Arun Kumar Gupta. "Diffusion weighted imaging: technique and applications." *World journal of radiology*, **8**(9):785, 2016.

[Ben19] E J Benjamin et al. "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association." *Circulation*, **139**:e56–e528, 11 2019.

[BFB15] Olvert A. Berkhemer, Puck S.S. Fransen, Debbie Beumer, Lucie A. van den Berg, Hester F. Lingsma, Albert J. Yoo, Wouter J. Schonewille, Jan Albert Vos, Paul J. Nederkoorn, Marieke J.H. Wermer, Marianne A.A. van Walderveen, Julie Staals, Jeannette Hofmeijer, Jacques A. van Oostayen, Geert J. Lycklama à Nijeholt, Jelis Boiten, Patrick A. Brouwer, Bart J. Emmer, Sebastiaan F. de Bruijn, Lukas C. van Dijk, L. Jaap Kappelle, Rob H. Lo, Ewoud J. van Dijk, Joost de Vries, Paul L.M. de Kort, Willem Jan J. van Rooij, Jan S.P. van den Berg, Boudewijn A.A.M. van Hasselt, Leo A.M. Aerden, René J. Dallinga, Marieke C. Visser, Joseph C.J. Bot, Patrick C. Vroomen, Omid Eshghi, Tobien H.C.M.L. Schreuder, Roel J.J. Heijboer, Koos Keizer, Alexander V. Tielbeek, Heleen M. den Hertog, Dick G. Gerrits, Renske M. van den Berg-Vos, Giorgos B. Karas, Ewout W. Steyerberg, H. Zwenneke Flach, Henk A. Marquering, Marieke E.S. Sprengers, Sjoerd F.M. Jenniskens, Ludo F.M. Beenen, René van den Berg, Peter J. Koudstaal, Wim H. van Zwam, Yvo B.W.E.M. Roos, Aad van der Lugt, Robert J. van Oostenbrugge, Charles B.L.M. Majoie, and Diederik W.J. Dippel. "A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke." *New England Journal of Medicine*, **372**(1):11–20, 2015. PMID: 25517348.

[BJB20] Chelsea A Boyd, Mahesh V Jayaraman, Grayson L Baird, William S Einhorn, Matthew T Stib, Michael K Atalay, Jerrold L Boxerman, Ana P Lourenco, Gaurav Jindal, Douglas T Hidlay, et al. "Detection of emergent large vessel occlusion stroke with CT angiography is high across all levels of radiology training and grayscale viewing methods." *European Radiology*, **30**:4447–4453, 2020.

[BLC09] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. "Curriculum Learning." pp. 41–48. Association for Computing Machinery, 2009.

[BLP18] Woong Bae, Seungho Lee, Gwangbeen Park, Hyunho Park, and Kyu-Hwan Jung. "Residual CNN-based image super-resolution for CT slice thickness reduction using paired CT scans: preliminary validation study." 2018.

[BMR20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." *Advances in neural information processing systems*, **33**:1877–1901, 2020.

[BRC17] Raphaël Blanc, Hocine Redjem, Gabriele Ciccio, Stanislas Smajda, Jean-Philippe Desilles, Eliane Orng, Guillaume Taylor, Elodie Drumez, Robert Fahed, Julien Labreuche, Mikael Mazighi, Bertrand Lapergue, and Michel Piotin. "Predictors of the Aspiration Component Success of a Direct Aspiration First Pass Technique (ADAPT) for the Endovascular Treatment of Stroke Reperfusion Strategy in Anterior Circulation Acute Stroke." *Stroke*, **48**:1588–1593, 6 2017.

[BSM19] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm." *Computers in biology and medicine*, **109**:218–225, 2019.

[BTL20] Romain Bourcier, Romain Thiaudière, Laurence Legrand, Benjamin Daumas-Duport, Hubert Desal, and Grégoire Boulouis. "Accelerated MR evaluation of patients with suspected large arterial vessel occlusion: diagnostic performances of the FLAIR vessel hyperintensities." *European Neurology*, **83**(4):389–394, 2020.

[BZL19] Qing-ke Bai, Zhen-guo Zhao, Lian-jun Lu, Jian Shen, Jian-ying Zhang, Hai-jing Sui, Xiu-hai Xie, Juan Chen, Juan Yang, and Cui-rong Chen. "Treating ischaemic stroke with intravenous tPA beyond 4.5 hours under the guidance of a MRI DWI/T2WI mismatch was safe and effective." *Stroke and Vascular Neurology*, **4**(1):8–13, 2019.

[CAL16] Ozgun Cicek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation." pp. 424–432, 2016.

[CBL17] Ángel Chamorro, Jordi Blasco, Antonio López, Sergio Amaro, Luis San Román, Laura Llull, Arturo Renú, Salvatore Rudilosso, Carlos Laredo, Victor Obach, et al. "Complete reperfusion is required for maximal benefits of mechanical thrombectomy in stroke patients." *Scientific reports*, **7**(1):11636, 2017.

[CBS22] Alexandra L Czap, Mersedeh Bahr-Hosseini, Noopur Singh, Jose-Miguel Yamal, May Nour, Stephanie Parker, Youngran Kim, Lucas Restrepo, Rania Abdelkhaleq, Sergio Salazar-Marioni, et al. "Machine learning automated detection of large vessel occlusion from mobile stroke unit computed tomography angiography." *Stroke*, **53**(5):1651–1656, 2022.

[CCC17] David Carrera, Bruce CV Campbell, Jordi Cortés, Montse Gorchs, Marisol Querol, Xavier Jiménez, Mònica Millán, Antoni Dávalos, and Natalia Pérez de la Ossa. "Predictive value of modifications of the prehospital rapid arterial occlusion evaluation scale for large vessel occlusion in patients with acute stroke." *Journal of Stroke and Cerebrovascular Diseases*, **26**(1):74–77, 2017.

[CFG20] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. "Improved baselines with momentum contrastive learning." *arXiv preprint arXiv:2003.04297*, 2020.

[CHC15] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition." *Plos One*, **10**, 2015.

[CHJ20] Min Chen, Nele Herregods, Jacob L Jaremko, Philippe Carron, Dirk Elewaut, Filip Van den Bosch, and Lennart Jans. "Diagnostic performance for erosion detection in sacroiliac joints on MR T1-weighted images: comparison between different slice thicknesses." *European Journal of Radiology*, **133**:109352, 2020.

[CKN07] Julio A Chalela, Chelsea S Kidwell, Lauren M Nentwich, Marie Luby, John A Butman, Andrew M Demchuk, Michael D Hill, Nicholas Patronas, Lawrence Latour, and Steven Warach. "Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison." *The Lancet*, **369**(9558):293–298, 2007.

[CKN20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[CMK15] Bruce C.V. Campbell, Peter J. Mitchell, Timothy J. Kleinig, Helen M. Dewey, Leonid Churilov, Nawaf Yassi, Bernard Yan, Richard J. Dowling, Mark W. Parsons, Thomas J. Oxley, Teddy Y. Wu, Mark Brooks, Marion A. Simpson, Ferdinand Miteff, Christopher R. Levi, Martin Krause, Timothy J. Harrington, Kenneth C. Faulder, Brendan S. Steinfort, Miriam Priglinger, Timothy Ang, Rebecca Scroop, P. Alan Barber, Ben McGuinness, Tissa Wijeratne, Thanh G.

Phan, Winston Chong, Ronil V. Chandra, Christopher F. Bladin, Monica Badve, Henry Rice, Laetitia de Villiers, Henry Ma, Patricia M. Desmond, Geoffrey A. Donnan, and Stephen M. Davis. "Endovascular Therapy for Ischemic Stroke with Perfusion-Imaging Selection." *New England Journal of Medicine*, **372**(11):1009–1018, 2015. PMID: 25671797.

[CSC18] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. "Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 91–99. Springer, 2018.

[CSH20] H P Chan, R K Samala, L M Hadjiiski, and C Zhou. "Deep Learning in Medical Image Analysis." *Adv. Exp. Med. Biol.*, **1213**:3–21, 2020.

[CTB19] Gustavo Carneiro, João Manuel R S Tavares, Andrew P Bradley, João Paulo Papa, Jacinto C Nascimento, Jaime S Cardoso, Zhi Lu, and Vasileios Belagiannis. "Editorial." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, **7**:241 – 241, 2019.

[CWG21] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. "Pre-trained image processing transformer." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12299–12310, 2021.

[DBK20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." 10 2020.

[DCB17] C. Dargazanli, A. Consoli, M. Barral, J. Labreuche, H. Redjem, G. Ciccio, S. Smajda, J.P. Desilles, G. Taylor, C. Preda, O. Coskun, G. Rodesch, M. Piotin, R. Blanc, and B. Lapergue. "Impact of Modified TICI 3 versus Modified TICI 2b Reperfusion Score to Predict Good Outcome following Endovascular Therapy." *American Journal of Neuroradiology*, **38**:90–96, 1 2017.

[DCL18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

[DKA16] Bart M. Demaerschalk, Dawn O. Kleindorfer, Opeolu M. Adeoye, Andrew M. Demchuk, Jennifer E. Fugate, James C. Grotta, Alexander A. Khalessi, Elad I. Levy, Yuko Y. Palesch, Shyam Prabhakaran, Gustavo Saposnik, Jeffrey L. Saver,

and Eric E. Smith. "Scientific Rationale for the Inclusion and Exclusion Criteria for Intravenous Alteplase in Acute Ischemic Stroke." *Stroke*, **47**, 2 2016.

[DLH15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Image super-resolution using deep convolutional networks." *IEEE transactions on pattern analysis and machine intelligence*, **38**(2):295–307, 2015.

[DPG20] Célina Ducroux, Michel Piotin, Benjamin Gory, Julien Labreuche, Raphael Blanc, Malek Ben Maacha, Bertrand Lapergue, and Robert Fahed. "First pass effect with contact aspiration and stent retrievers in the Aspiration versus Stent Retriever (ASTER) trial." *Journal of NeuroInterventional Surgery*, **12**:386–391, 4 2020.

[dTP17] Christopher D d'Esterre, Anurag Trivedi, Pooneh Pordeli, Mari Boesen, Shivanand Patil, Seong Hwan Ahn, Mohamed Najm, Enrico Fainardi, Jai Jai Shiva Shankar, Marta Rubiera, et al. "Regional comparison of multiphase computed tomographic angiography and computed tomographic perfusion for prediction of tissue fate in ischemic stroke." *Stroke*, **48**(4):939–945, 2017.

[EBS18] Mark R. Etherton, Andrew D. Barreto, Lee H. Schwamm, and Ona Wu. "Neuroimaging Paradigms to Identify Patients for Reperfusion Therapy in Stroke of Unknown Onset." *Frontiers in Neurology*, **9**, 5 2018.

[Elm93] Jeffrey L Elman. "Learning and development in neural networks: the importance of starting small." *Cognition*, **48**:71 – 99, 1993.

[ESN13] Marco Essig, Mark S Shiroishi, Thanh Binh Nguyen, Marc Saake, James M Provenzale, David Enterline, Nicoletta Anzalone, Arnd Dörfler, Àlex Rovira, Max Wintermark, et al. "Perfusion MRI: the five most frequently asked technical questions." *AJR. American journal of roentgenology*, **200**(1):24, 2013.

[FBF21] Fabian Flottmann, Gabriel Broocks, Tobias Djamsched Faizy, Rosalie McDonough, Lucas Watermann, Milani Deb-Chatterji, Götz Thomalla, Moriz Herzberg, Christian H. Nolte, Jens Fiehler, Hannes Leischner, and Caspar Brekenfeld. "Factors Associated with Failure of Reperfusion in Endovascular Therapy for Acute Ischemic Stroke." *Clinical Neuroradiology*, **31**:197–205, 3 2021.

[FBH21] Bram ACM Fasen, Rob AP Borghans, Roeland JJ Heijboer, Frans-Jan H Hulsmans, and Robert M Kwee. "Reliability and accuracy of 3-mm and 2-mm maximum intensity projection CT angiography to detect intracranial large vessel occlusion in patients with acute anterior cerebral circulation stroke." *Neuroradiology*, pp. 1–6, 2021.

[FEM09] Vladimir Fonov, Alan Evans, Robert Mckinstry, C R Almli, and Louis Collins. "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood." *Neuroimage*, **47**, 11 2009.

[FKK13] J.E. Fugate, A.M. Klunder, and D.F. Kallmes. "What Is Meant by "TICI"?" *American Journal of Neuroradiology*, **34**:1792–1797, 9 2013.

[FS16] Andrew A. Fanous and Adnan H. Siddiqui. "Mechanical thrombectomy: Stent retrievers vs. aspiration catheters." *Cor et Vasa*, **58**(2):e193–e203, 2016. Acute Ischemic Stroke.

[Fuk80] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." *Biological cybernetics*, **36**(4):193–202, 1980.

[GDM15] Mayank Goyal, Andrew M. Demchuk, Bijoy K. Menon, Muneer Eesa, Jeremy L. Rempel, John Thornton, Daniel Roy, Tudor G. Jovin, Robert A. Willinsky, Biggya L. Sapkota, Dar Dowlatshahi, Donald F. Frei, Noreen R. Kamal, Walter J. Montanera, Alexandre Y. Poppe, Karla J. Ryckborst, Frank L. Silver, Ashfaq Shuaib, Donatella Tampieri, David Williams, Oh Young Bang, Blaise W. Baxter, Paul A. Burns, Hana Choe, Ji-Hoe Heo, Christine A. Holmstedt, Brian Jankowitz, Michael Kelly, Guillermo Linares, Jennifer L. Mandzia, Jai Shankar, Sung-Il Sohn, Richard H. Swartz, Philip A. Barber, Shelagh B. Coutts, Eric E. Smith, William F. Morrish, Alain Weill, Suresh Subramaniam, Alim P. Mitha, John H. Wong, Mark W. Lowerison, Tolulope T. Sajobi, and Michael D. Hill. "Randomized Assessment of Rapid Endovascular Treatment of Ischemic Stroke." *New England Journal of Medicine*, **372**(11):1019–1030, 2015. PMID: 25671798.

[GGB20] Aglaé Velasco Gonzalez, Dennis Görlich, Boris Buerke, Nico Münnich, Cristina Sauerland, Thilo Rusche, Andreas Faldum, and Walter Heindel. "Predictors of Successful First-Pass Thrombectomy with a Balloon Guide Catheter: Results of a Decision Tree Analysis." *Translational Stroke Research*, **11**:900–909, 10 2020.

[GMV16] Mayank Goyal, Bijoy K Menon, Wim H Van Zwam, Diederik WJ Dippel, Peter J Mitchell, Andrew M Demchuk, Antoni Dávalos, Charles BLM Majoie, Aad van Der Lugt, Maria A De Miquel, et al. "Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials." *The Lancet*, **387**(10029):1723–1731, 2016.

[Gor19] Philip B Gorelick. "The global burden of stroke: persistent and disabling." *The Lancet Neurology*, **18**:417–418, 5 2019.

[GSA20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. "Bootstrap your own latent-a new approach to self-supervised learning." *Advances in neural information processing systems*, **33**:21271–21284, 2020.

[GTF19] Nitin Goyal, Georgios Tsivgoulis, Donald Frei, Aquilla Turk, Blaise Baxter, Michael T Froehler, J Mocco, Muhammad Fawad Ishfaq, Konark Malhotra, Jason J Chang, Daniel Hoit, Lucas Elijovich, David Loy, Raymond D Turner, Justin Mascitelli, Kiersten Espaillat, Andrei V Alexandrov, and Adam S Arthur. "Comparative Safety and Efficacy of Modified TICI 2b and TICI 3 Reperfusion in Acute Ischemic Strokes Treated With Mechanical Thrombectomy." *Neurosurgery*, **84**:680–686, 3 2019.

[GXL22] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. "Attention mechanisms in computer vision: A survey." *Computational Visual Media*, **8**(3):331–368, 2022.

[GYX19] Rongjun Ge, Guanyu Yang, Chenchu Xu, Yang Chen, Limin Luo, and Shuo Li. "Stereo-correlation and noise-distribution aware ResVoxGAN for dense slices reconstruction and noise reduction in thick low-dose CT." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pp. 328–338. Springer, 2019.

[HBK92] Joseph V Hajnal, David J Bryant, Larry Kasuboski, Pradip M Pattany, Beatrice De Coene, Paul D Lewis, Jacqueline M Pennock, Angela Oatridge, Ian R Young, and Graeme M Bydder. "Use of fluid attenuated inversion recovery (FLAIR) pulse sequences in MRI of the brain." *Journal of computer assisted tomography*, **16**(6):841–844, 1992.

[HBR20] Jeremy Hofmeister, Gianmarco Bernava, Andrea Rosi, Maria Isabel Vargas, Emmanuel Carrera, Xavier Montet, Simon Burgermeister, Pierre-Alexandre Poletti, Alexandra Platon, Karl-Olof Lovblad, and Paolo Machi. "Clot-Based Radiomics Predict a Mechanical Thrombectomy Strategy for Successful Recanalization in Acute Ischemic Stroke." *Stroke*, **51**:2488–2494, 8 2020.

[HCX21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked Autoencoders Are Scalable Vision Learners." 11 2021.

[HF03] Randall T Higashida and Anthony J Furlan. "Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke." *stroke*, **34**(8):e109–e137, 2003.

[HFW20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[HHB16] Mirjam R Heldner, Kety Hsieh, Anne Broeg-Morvay, Pasquale Mordasini, Monika Bühlmann, Simon Jung, Marcel Arnold, Heinrich P Mattle, Jan Gralla, and Urs Fischer. "Clinical prediction of large vessel occlusion in anterior circulation stroke: mission impossible?" *Journal of neurology*, **263**:1633–1640, 2016.

[HHJ21] Yung-Pin Hwang, Chun-Chao Huang, Zong-Yi Jhou, Wei-Ming Huang, Helen L Po, and Chao-Liang Chou. "Using Glasgow coma scale to identify acute large-vessel occlusion stroke." *International Journal of Gerontology*, **15**(1):64–66, 2021.

[HHM16] Lan He, Yanqi Huang, Zelan Ma, Cuishan Liang, Changhong Liang, and Zaiyi Liu. "Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule." *Scientific reports*, **6**(1):34921, 2016.

[HKB08] Werner Hacke, Markku Kaste, Erich Bluhmki, Miroslav Brozman, Antoni Dávalos, Donata Guidetti, Vincent Larrue, Kennedy R. Lees, Zakaria Medeghri, Thomas Machnig, Dietmar Schneider, Rüdiger von Kummer, Nils Wahlgren, and Danilo Toni. "Thrombolysis with Alteplase 3 to 4.5 Hours after Acute Ischemic Stroke." *New England Journal of Medicine*, **359**:1317–1329, 9 2008.

[HKM22] P Anirudh Hebbar, MV Manoj Kumar, and Archana Mathur. "Theory, concepts, and applications of artificial neural networks." In *Applied Soft Computing*, pp. 153–176. Apple Academic Press, 2022.

[HLV17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[HRO19] A. Hilbert, L.A. Ramos, H.J.A. van Os, S.D. Olabarriaga, M.L. Tolhuisen, M.J.H. Wermer, R.S. Barros, I. van der Schaaf, D. Dippel, Y.B.W.E.M. Roos, W.H. van Zwam, A.J. Yoo, B.J. Emmer, G.J. Lycklama à Nijeholt, A.H. Zwinderman, G.J. Strijkers, C.B.L.M. Majoie, and H.A. Marquering. "Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke." *Computers in Biology and Medicine*, **115**:103516, 12 2019.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." *Neural computation*, **9**(8):1735–1780, 1997.

[HSE17] K C Ho, W Speier, S El-Saden, and C W Arnold. "Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features." *AMIA Annu Symp Proc*, **2017**:892–901, 2017.

[HSS18] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[HSZ19] King Chung Ho, William Speier, Haoyue Zhang, Fabien Scalzo, Suzie El-Saden, and Corey W. Arnold. "A Machine Learning Approach for Classifying Ischemic Stroke Onset Time From Imaging." *IEEE Transactions on Medical Imaging*, **38**:1666–1676, 7 2019.

[HTC13] Christine A Holmstedt, Tanya N Turan, and Marc I Chimowitz. "Atherosclerotic intracranial arterial stenosis: risk factors, diagnosis, and treatment." *The Lancet Neurology*, **12**(11):1106–1114, 2013.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[IJK21] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature methods*, **18**(2):203–211, 2021.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

[Jaf19] Mahdad Jafarzadeh Esfahani. "Minimal sensing of quality of upper-limb movements." August 2019.

[JCC15] Tudor G. Jovin, Angel Chamorro, Erik Cobo, María A. de Miquel, Carlos A. Molina, Alex Rovira, Luis San Román, Joaquín Serena, Sonia Abilleira, Marc Ribó, Mònica Millán, Xabier Urra, Pere Cardona, Elena López-Cancio, Alejandro Tomasello, Carlos Castaño, Jordi Blasco, Lucía Aja, Laura Dorado, Helena Quesada, Marta Rubiera, María Hernandez-Pérez, Mayank Goyal, Andrew M. Demchuk, Rüdiger von Kummer, Miquel Gallofré, and Antoni Dávalos. "Thrombectomy within 8 Hours after Symptom Onset in Ischemic Stroke." *New England Journal of Medicine*, **372**(24):2296–2306, 2015. PMID: 25882510.

[JCW19] Gaurav Jindal, Helio De Paula Carvalho, Aaron Wessell, Elizabeth Le, Varun Naragum, Timothy Ryan Miller, Marcella Wozniak, Ravi Shivashankar, Carolyn A Cronin, Chad Schrier, and Dheeraj Gandhi. "Beyond the first pass: revascularization remains critical in stroke thrombectomy." *Journal of NeuroInterventional Surgery*, **11**:1095–1099, 11 2019.

[JWG16] Heesoo Joo, Guijing Wang, and Mary G George. "Use of intravenous tissue plasminogen activator and hospital costs for patients with acute ischaemic stroke aged 18–64 years in the USA." *BMJ*, **1**, 3 2016.

[KBW19] Jayme C. Kosior, Brian Buck, Robert Wannamaker, Mahesh Kate, Natalia A. Liapounova, Jeremy L. Rempel, and Kenneth Butcher. "Exploring Reperfusion Following Endovascular Thrombectomy." *Stroke*, **50**:2389–2395, 9 2019.

[KFW03] Fumiko Kodama, Patrick J Fultz, and John C Wandtke. "Comparing thin-section and thick-section CT of pericardial sinuses and recesses." *American Journal of Roentgenology*, **181**(4):1101–1108, 2003.

[KLL16] 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*, pp. 1646–1654, 2016.

[KNH21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. "Transformers in vision: A survey." *arXiv preprint arXiv:2101.01169*, 2021.

[KP06] Dow-Muh Koh and Anwar R Padhani. "Diffusion-weighted MRI: a new functional clinical technique for tumour imaging." *The British Journal of Radiology*, **79**(944):633–635, 2006.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM*, **60**(6):84–90, 2017.

[Lai14] Vincent Lai. "Application of diffusion–and perfusion–weighted imaging in acute ischemic stroke." In *Advanced Brain Neuroimaging Topics in Health and Disease-Methods and Applications*. IntechOpen, 2014.

[LBD89] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. "Backpropagation applied to handwritten zip code recognition." *Neural computation*, **1**(4):541–551, 1989.

[LBG17] Bertrand Lapergue, Raphael Blanc, Benjamin Gory, Julien Labreuche, Alain Duhamel, Gautier Marnat, Suzana Saleme, Vincent Costalat, Serge Bracard, Hubert Desal, Mikael Mazighi, Arturo Consoli, and Michel Piotin. "Effect of Endovascular Contact Aspiration vs Stent Retriever on Revascularization in Patients With Acute Ischemic Stroke and Large Vessel Occlusion." *JAMA*, **318**:443, 8 2017.

[LBK10] Kennedy R Lees, Erich Bluhmki, Rüdiger von Kummer, Thomas G Brott, Danilo Toni, James C Grotta, Gregory W Albers, Markku Kaste, John R Marler, Scott A Hamilton, Barbara C Tilley, Stephen M Davis, Geoffrey A Donnan, and Werner Hacke. "Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials." *The Lancet*, **375**:1695–1703, 5 2010.

[LCS21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. "Swinir: Image restoration using swin transformer." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021.

[LCX21] Gaoping Liu, Zehong Cao, Qiang Xu, Qirui Zhang, Fang Yang, Xinyu Xie, Jingru Hao, Yinghuan Shi, Boris C Bernhardt, Yichu He, et al. "Recycling diagnostic MRI for empowering brain morphometric research–Critical & practical assessment on learning-based image super-resolution." *Neuroimage*, **245**:118687, 2021.

[LFH19] Hannes Leischner, Fabian Flottmann, Uta Hanning, Gabriel Broocks, Tobias Djamsched Faizy, Milani Deb-Chatterji, Martina Bernhardt, Caspar Brekenfeld, Jan-Hendrik Buhk, Susanne Gellissen, Götz Thomalla, Christian Gerloff, and Jens Fiehler. "Reasons for failed endovascular recanalization attempts in stroke patients." *Journal of NeuroInterventional Surgery*, **11**:439–442, 5 2019.

[Lit17] Geert Litjens et al. "A survey on deep learning in medical image analysis." *Medical Image Analysis*, **42**:60 – 88, 2017.

[LLC21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." *arXiv preprint arXiv:2103.14030*, 2021.

[LLH20] Hyunna Lee, Eun-Jae Lee, Sungwon Ham, Han-Bin Lee, Ji Sung Lee, Sun U. Kwon, Jong S. Kim, Namkug Kim, and Dong-Wha Kang. "Machine Learning Approach to Identify Stroke Within 4.5 Hours." *Stroke*, **51**:860–866, 3 2020.

[LLR21] Su Jin Lee, Belinda Liu, Neil Rane, Peter Mitchell, Richard Dowling, and Bernard Yan. "Correlation between CT angiography and digital subtraction angiography in acute ischemic strokes." *Clinical Neurology and Neurosurgery*, **200**:106399, 2021.

[LLW21] Zhiyang Lu, Zheng Li, Jun Wang, Jun Shi, and Dinggang Shen. "Two-stage self-supervised cycle-consistency network for reconstruction of thin-slice mr images." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pp. 3–12. Springer, 2021.

[LMW22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A convnet for the 2020s." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

[LSD20] Benjamin J Lawner, Kelly Szabo, Jonathan Daly, Krista Foster, Philip McCoy, David Poliner, Matthew Poremba, Philip S Nawrocki, and Rahul Rahangdale. "Challenges related to the implementation of an EMS-administered, large vessel occlusion stroke score." *Western Journal of Emergency Medicine*, **21**(2):441, 2020.

[LSK17] 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, pp. 136–144, 2017.

[LTH17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan 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*, pp. 4681–4690, 2017.

[LXL19] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. "Adaptive Gradient Methods with Dynamic Bound of Learning Rate." 5 2019.

[LZL20] Qiuyue Liu, Zhen Zhou, Feng Liu, Xiangming Fang, Yizhou Yu, and Yizhou Wang. "Multi-stream progressive up-sampling network for dense CT image reconstruction." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pp. 518–528. Springer, 2020.

[MCP19] Henry Ma, Bruce C.V. Campbell, Mark W. Parsons, Leonid Churilov, Christopher R. Levi, Chung Hsu, Timothy J. Kleinig, Tissa Wijeratne, Sami Curtze, Helen M. Dewey, Ferdinand Miteff, Chon-Haw Tsai, Jiunn-Tay Lee, Thanh G. Phan, Neil Mahant, Mu-Chien Sun, Martin Krause, Jonathan Sturm, Rohan Grimley, Chih-Hung Chen, Chaur-Jong Hu, Andrew A. Wong, Deborah Field, Yu Sun, P. Alan Barber, Arman Sabet, Jim Jannes, Jiann-Shing Jeng, Benjamin Clissold, Romesh Markus, Ching-Huang Lin, Li-Ming Lien, Christopher F. Bladin, Søren Christensen, Nawaf Yassi, Gagan Sharma, Andrew Bivard, Patricia M. Desmond, Bernard Yan, Peter J. Mitchell, Vincent Thijs, Leeanne Carey, Atte Meretoja, Stephen M. Davis, and Geoffrey A. Donnan. "Thrombolysis Guided by Perfusion Imaging up to 9 Hours after Onset of Stroke." *New England Journal of Medicine*, **380**:1795–1803, 5 2019.

[MHG14] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention." *Advances in neural information processing systems*, **27**, 2014.

[MHO20] Ryan A. McTaggart, Jessalyn K. Holodinsky, Johanna M. Ospel, Andrew K. Cheung, Nathan W. Manning, Jason D. Wenderoth, Thanh G. Phan, Richard Beare, Kendall Lane, Richard A. Haas, Noreen Kamal, Mayank Goyal, and Mahesh V. Jayaraman. "Leaving No Large Vessel Occlusion Stroke Behind." *Stroke*, **51**:1951–1960, 7 2020.

[MKC21] Federico Di Maria, Maéva Kyheng, Arturo Consoli, Jean-Philippe Desilles, Benjamin Gory, Sébastien Richard, Georges Rodesch, Julien Labreuche, Jean-Baptiste Girot, Cyril Dargazanli, Gaultier Marnat, Bertrand Lapergue, and Romain Bourcier. "Identifying the predictors of first-pass effect and its influence on

clinical outcome in the setting of endovascular thrombectomy for acute ischemic stroke: Results from a multicentric prospective registry." *International Journal of Stroke*, **16**:20–28, 1 2021.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

[NHL21] Raul G. Nogueira, Diogo C. Haussen, David Liebeskind, Tudor G. Jovin, Rishi Gupta, Ashutov Jadhav, Ron F. Budzik, Blaise Baxter, Antonin Krajina, Alain Bonafe, Ali Malek, Ana Paula Narata, Ryan Shields, Yanchang Zhang, Patricia Morgan, Bruno Bartolini, Joey English, Michael R. Frankel, and Erol Veznedaroglu. "Stroke Imaging Selection Modality and Endovascular Therapy Outcomes in the Early and Extended Time Windows." *Stroke*, **52**:491–497, 2 2021.

[NIM22] Jennifer K Nicholls, Jonathan Ince, Jatinder S Minhas, and Emma ML Chung. "Emerging detection techniques for large vessel occlusion stroke: a scoping review." *Frontiers in Neurology*, **12**:2477, 2022.

[NJH18] Raul G. Nogueira, Ashutosh P. Jadhav, Diogo C. Haussen, Alain Bonafe, Ronald F. Budzik, Parita Bhuva, Dileep R. Yavagal, Marc Ribo, Christophe Cognard, Ricardo A. Hanel, Cathy A. Sila, Ameer E. Hassan, Monica Millan, Elad I. Levy, Peter Mitchell, Michael Chen, Joey D. English, Qaisar A. Shah, Frank L. Silver, Vitor M. Pereira, Brijesh P. Mehta, Blaise W. Baxter, Michael G. Abraham, Pedro Cardona, Erol Veznedaroglu, Frank R. Hellinger, Lei Feng, Jawad F. Kirmani, Demetrius K. Lopes, Brian T. Jankowitz, Michael R. Frankel, Vincent Costalat, Nirav A. Vora, Albert J. Yoo, Amer M. Malik, Anthony J. Furlan, Marta Rubiera, Amin Aghaebrahim, Jean-Marc Olivot, Wondwossen G. Tekle, Ryan Shields, Todd Graves, Roger J. Lewis, Wade S. Smith, David S. Liebeskind, Jeffrey L. Saver, and Tudor G. Jovin. "Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct." *New England Journal of Medicine*, **378**:11–21, 1 2018.

[NSS18] Ali Reza Noorian, Nerses Sanossian, Kristina Shkirkova, David S Liebeskind, Marc Eckstein, Samuel J Stratton, Franklin D Pratt, Robin Conwit, Fiona Chatfield, Latisha K Sharma, et al. "Los Angeles Motor Scale to identify large vessel occlusion: prehospital validation and comparison with other screens." *Stroke*, **49**(3):565–572, 2018.

[OCG20] Marta Olive-Gadea, Carlos Crespo, Cristina Granes, Maria Hernandez-Perez, Natalia Perez de la Ossa, Carlos Laredo, Xabier Urra, Juan Carlos Soler, Alexander Soler, Paloma Puyalto, et al. "Deep learning based software to identify large vessel occlusion on noncontrast computed tomography." *Stroke*, **51**(10):3133–3137, 2020.

[OSF18] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C H Lee, Mattias P Heinrich, Kazunari Misawa, Kensaku Mori, Steven G McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention U-Net: Learning Where to Look for the Pancreas." *ArXiv*, **abs/1804.03999**, 2018.

[PHR17] Jan Christoph Purrucker, Florian Härtig, Hardy Richter, Andreas Engelbrecht, Johannes Hartmann, Jonas Auer, Christian Hametner, Erik Popp, Peter Arthur Ringleb, Simon Nagel, et al. "Design and validation of a clinical scale for prehospital stroke recognition, severity grading and prediction of large vessel occlusion: the shortened NIH Stroke Scale for emergency medical services." *bmj Open*, **7**(9):e016893, 2017.

[PK12] Donald B Plewes and Walter Kucharczyk. "Physics of MRI: a primer." *Journal of magnetic resonance imaging*, **35**(5):1038–1054, 2012.

[PLK21] Sohee Park, Sang Min Lee, Wooil Kim, Hyunho Park, Kyu-Hwan Jung, Kyung-Hyun Do, and Joon Beom Seo. "Computer-aided detection of subsolid nodules at chest CT: improved performance with deep learning–based CT section thickness reduction." *Radiology*, **299**(1):211–219, 2021.

[PLL20] Cheng Peng, Wei-An Lin, Haofu Liao, Rama Chellappa, and S Kevin Zhou. "Saint: spatially aware interpolation network for medical slice synthesis." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7750–7759, 2020.

[PRA19] William J. Powers, Alejandro A. Rabinstein, Teri Ackerson, Opeolu M. Adeoye, Nicholas C. Bambakidis, Kyra Becker, José Biller, Michael Brown, Bart M. Demaerschalk, Brian Hoh, Edward C. Jauch, Chelsea S. Kidwell, Thabele M. Leslie-Mazwi, Bruce Ovbiagele, Phillip A. Scott, Kevin N. Sheth, Andrew M. Southerland, Deborah V. Summers, and David L. Tirschwell. "Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association." *Stroke*, **50**, 12 2019.

[PSG19] Guilherme Santos Piedade, Clemens M. Schirmer, Oded Goren, Hua Zhang, Amir Aghajanian, James E. Faber, and Christoph J. Griessenauer. "Cerebral Collateral Circulation: A Review in the Context of Ischemic Stroke and Mechanical Thrombectomy." *World Neurosurgery*, **122**:33–42, 2 2019.

[PSL19] Corentin Provost, Marc Soudant, Laurence Legrand, Wagih Ben Hassen, Yu Xie, Sébastien Soize, Romain Bourcier, Joseph Benzakoun, Myriam Edjlali, Grégoire Boulouis, et al. "Magnetic resonance imaging or computed tomography before treatment in acute ischemic stroke: effect on workflow and functional outcome." *Stroke*, **50**(3):659–664, 2019.

[PVG19] Christopher A Potter, Achala S Vagal, Mayank Goyal, Diego B Nunez, Thabele M Leslie-Mazwi, and Michael H Lev. "CT for treatment selection in acute ischemic stroke: a code stroke primer." *Radiographics*, **39**(6):1717–1738, 2019.

[PY10] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering*, **22**:1345–1359, 2010.

[PZC21] Cheng Peng, S Kevin Zhou, and Rama Chellappa. "DA-VSR: domain adaptable volumetric super-resolution for medical images." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pp. 75–85. Springer, 2021.

[PZN21] Jennifer Polson, Haoyue Zhang, Kambiz Nael, Noriko Salamon, Bryan Yoo, Namkug Kim, Dong-Wha Kang, William Speier, and Corey W. Arnold. "A Semi-Supervised Learning Framework to Leverage Proxy Information for Stroke MRI Analysis." pp. 2258–2261. IEEE, 11 2021.

[PZN22] Jennifer S. Polson, Haoyue Zhang, Kambiz Nael, Noriko Salamon, Bryan Y. Yoo, Suzie El-Saden, Sidney Starkman, Namkug Kim, Dong-Wha Kang, William F. Speier, and Corey W. Arnold. "Identifying acute ischemic stroke patients within the thrombolytic treatment window using deep learning." *Journal of Neuroimaging*, **32**:1153–1160, 11 2022.

[QKN19] W. Qiu, H. Kuang, J. Nair, Z. Assis, M. Najm, C. McDougall, B. McDougall, K. Chung, A.T. Wilson, M. Goyal, M.D. Hill, A.M. Demchuk, and B.K. Menon. "Radiomics-Based Intracranial Thrombus Features on CT and CTA Predict Recanalization with Intravenous Alteplase in Patients with Acute Ischemic Stroke." *American Journal of Neuroradiology*, **40**:39–44, 1 2019.

[RBT14] Vladimir Rohan, Jan Baxa, Radek Tupy, Lenka Cerna, Petr Sevcik, Michal Friesl, Jiri Polivka, Jiri Polivka, and Jiri Ferda. "Length of Occlusion Predicts Recanalization and Outcome After Intravenous Thrombolysis in Middle Cerebral Artery Stroke." *Stroke*, **45**:2010–2017, 7 2014.

[RDS15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S Bernstein, Alexander C Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision*, **115**:211–252, 2015.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." 5 2015.

[RHT18] Christopher T Richards, Ryan Huebinger, Katie L Tataris, Joseph M Weber, Laura Eggers, Eddie Markul, Leslee Stein-Spencer, Kenneth S Pearlman, Jane L

Holl, and Shyam Prabhakaran. "Cincinnati prehospital stroke scale can identify large vessel occlusion stroke." *Prehospital Emergency Care*, **22**(3):312–318, 2018.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors." *nature*, **323**(6088):533–536, 1986.

[RSS02] P.A. Ringleb, P.D. Schellinger, C. Schranz, and W. Hacke. "Thrombolytic Therapy Within 3 to 6 Hours After Onset of Ischemic Stroke." *Stroke*, **33**:1437–1441, 5 2002.

[RT14] D. Leander Rimmele and GÃ¶tz Thomalla. "Wake-Up Stroke: Clinical Characteristics, Imaging Findings, and Treatment Option â€“ an Update." *Frontiers in Neurology*, **5**, 3 2014.

[RWS19] Robert C Rennert, Arvin R Wali, Jeffrey A Steinberg, David R Santiago-Dieppa, Scott E Olson, J Scott Pannell, and Alexander A Khalessi. "Epidemiology, Natural History, and Clinical Presentation of Large Vessel Ischemic Stroke." *Neurosurgery*, **85**:S4–S8, 7 2019.

[RZK19] Maithra Raghu, Chiyuan Zhang, Jon M Kleinberg, and Samy Bengio. "Transfusion: Understanding Transfer Learning for Medical Imaging." 2019.

[SBM11] Ashfaq Shuaib, Ken Butcher, Askar A Mohammad, Maher Saqqur, and David S Liebeskind. "Collateral blood vessels in acute ischaemic stroke: a potential therapeutic target." *The Lancet Neurology*, **10**:909–921, 10 2011.

[Sca13] F Scalzo et al. "Multi-center prediction of hemorrhagic transformation in acute ischemic stroke using permeability imaging features." *Magn Reson Imaging*, **31**:961–969, 7 2013.

[SCD19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *International Journal of Computer Vision*, **128**:336–359, 10 2019.

[SCH16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

[SDS20] Shaarada Srivatsa, Yifei Duan, John P. Sheppard, Shivani Pahwa, Jonathan Pace, Xiaofei Zhou, and Nicholas C. Bambakidis. "Cerebral vessel anatomy as a predictor of first-pass effect in mechanical thrombectomy for emergent large-vessel occlusion." *Journal of Neurosurgery*, **134**:576–584, 1 2020.

[SGL16] Jeffrey L Saver, Mayank Goyal, AAD Van der Lugt, Bijoy K Menon, Charles BLM Majoie, Diederik W Dippel, Bruce C Campbell, Raul G Nogueira, Andrew M Demchuk, Alejandro Tomasello, et al. "Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis." *Jama*, **316**(12):1279–1289, 2016.

[Shi16] H C Shin et al. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning." *IEEE Trans Med Imaging*, **35**:1285–1298, 11 2016.

[SHK14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research*, **15**(1):1929–1958, 2014.

[SJW04] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E.J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. "Advances in functional and structural MR image analysis and implementation as FSL." *NeuroImage*, **23**:S208–S219, 1 2004.

[SMS99] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. "Policy gradient methods for reinforcement learning with function approximation." *Advances in neural information processing systems*, **12**, 1999.

[SOS19] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention gated networks: Learning to leverage salient regions in medical images." *Medical Image Analysis*, **53**, 2019.

[SWS18] Lee H. Schwamm, Ona Wu, Shlee S. Song, Lawrence L. Latour, Andria L. Ford, Amie W. Hsia, Alona Muzikansky, Rebecca A. Betensky, Albert J. Yoo, Michael H. Lev, Gregoire Boulouis, Arne Lauer, Pedro Cougo, William A. Copen, Gordon J. Harris, Steven Warach, Sidney Starkman, Ramin Zand, Kendra Drake, Carlos Kase, Raphael Carandang, and Eric Searls. "Intravenous thrombolysis in unwitnessed stroke onset: MR WITNESS trial results." *Annals of Neurology*, **83**:980–993, 5 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*, 2014.

[TAC10] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. "N4ITK: Improved N3 Bias Correction." *IEEE Transactions on Medical Imaging*, **29**:1310–1320, 2010.

[TFO14] G Thomalla, Jochen B. Fiebach, Leif Ostergaard, Salvador Pedraza, Vincent Thijs, Norbert Nighoghossian, Pascal Roy, Keith W. Muir, Martin Ebinger, Bastian Cheng, Ivana Galinovic, Tae-Hee Cho, Josep Puig, Florent Boutitie, Claus Z. Simonsen, Matthias Endres, Jens Fiehler, and Christian Gerloff. "A Multicenter, Randomized, Double-Blind, Placebo-Controlled Trial to Test Efficacy and Safety of Magnetic Resonance Imaging-Based Thrombolysis in Wake-up Stroke (WAKE-UP)." *International Journal of Stroke*, **9**, 8 2014.

[Tho11] G Thomalla et al. "DWI-FLAIR mismatch for the identification of patients with acute ischaemic stroke within 4Â·5 h of symptom onset (PRE-FLAIR): a multicentre observational study." *Lancet Neurol*, **10**:978–986, 11 2011.

[TL19] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

[Tom07] Thomas Tomsick. "TIMI, TIBI, TICI: I came, I saw, I got confused." *AJNR. American journal of neuroradiology*, **28**:382–4, 2 2007.

[TSB18] Götz Thomalla, Claus Z. Simonsen, Florent Boutitie, Grethe Andersen, Yves Berthezene, Bastian Cheng, Bharath Cheripelli, Tae-Hee Cho, Franz Fazekas, Jens Fiehler, Ian Ford, Ivana Galinovic, Susanne Gellissen, Amir Golsari, Johannes Gregori, Matthias Günther, Jorge Guibernau, Karl Georg Häusler, Michael Hennerici, André Kemmling, Jacob Marstrand, Boris Modrau, Lars Neeb, Natalia Perez de la Ossa, Josep Puig, Peter Ringleb, Pascal Roy, Enno Scheel, Wouter Schonewille, Joaquin Serena, Stefan Sunaert, Kersten Villringer, Anke Wouters, Vincent Thijs, Martin Ebinger, Matthias Endres, Jochen B. Fiebach, Robin Lemmens, Keith W. Muir, Norbert Nighoghossian, Salvador Pedraza, and Christian Gerloff. "MRI-Guided Thrombolysis for Stroke with Unknown Time of Onset." *New England Journal of Medicine*, **379**:611–622, 8 2018.

[TSS19] Muhammad Asif Taqi, Ajeet Sodhi, Sajid S Suriya, Syed A Quadri, Mudassir Farooqui, Angelo A Salvucci, Adriane Stefansen, Martin M Mortazavi, and Daniel Shepherd. "Design, application and infield validation of a pre-hospital emergent large vessel occlusion screening tool: ventura emergent large vessel occlusion score." *Journal of Stroke and Cerebrovascular Diseases*, **28**(3):728–734, 2019.

[TVC17] Mohamed S Teleb, Anna Ver Hage, Jaqueline Carter, Mahesh V Jayaraman, and Ryan A McTaggart. "Stroke vision, aphasia, neglect (VAN) assessment—a novel emergent large vessel occlusion screening tool: pilot study and comparison with current clinical severity indices." *Journal of neurointerventional surgery*, **9**(2):122–126, 2017.

[TZY22] Wen Tang, Haoyue Zhang, Pengxin Yu, Han Kang, and Rongguo Zhang. "MMMNA-Net for Overall Survival Time Prediction of Brain Tumor Patients."

In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3805–3808. IEEE, 2022.

[UFZ18] Victor C. Urrutia, Roland Faigle, Steven R. Zeiler, Elisabeth B. Marsh, Mona Bahouth, Mario Cerdan Trevino, Jennifer Dearborn, Richard Leigh, Susan Rice, Karen Lane, Mustapha Saheed, Peter Hill, and Rafael H. Llinas. "Safety of intravenous alteplase within 4.5 hours for patients awakening with stroke symptoms." *PLOS ONE*, **13**:e0197714, 5 2018.

[VAB20] Salim S Virani, Alvaro Alonso, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Alexander R Chang, Susan Cheng, Francesca N Delling, et al. "Heart disease and stroke statistics—2020 update: a report from the American Heart Association." *Circulation*, **141**(9):e139–e596, 2020.

[VAF19] Simone Vidale, Marco Arnaboldi, Lara Frangi, Marco Longoni, Gianmario Monza, and Elio Agostoni. "The large artery intracranial occlusion stroke scale: A new tool with high accuracy in predicting large vessel occlusion." *Frontiers in Neurology*, **10**:130, 2019.

[VCS17] Ondrej Volny, Petra Cimflova, and Viktor Szeder. "Inter-Rater Reliability for Thrombolysis in Cerebral Infarction with TICI 2c Category." *Journal of Stroke and Cerebrovascular Diseases*, **26**:992–994, 5 2017.

[VMS21] Lohit Velagapudi, Nikolaos Mouchtouris, Richard F. Schmidt, David Vuong, Omaditya Khanna, Ahmad Sweid, Bryan Sadler, Fadi Al Saiegh, M. Reid Gooch, Pascal Jabbour, Robert H. Rosenwasser, and Stavropoula Tjoumakaris. "A Machine Learning Approach to First Pass Reperfusion in Mechanical Thrombectomy: Prediction and Feature Analysis." *Journal of Stroke and Cerebrovascular Diseases*, **30**:105796, 7 2021.

[VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." 6 2017.

[VVM20] Edward Verenich, Alvaro Velasquez, M G Sarwar Murshed, and Faraz Hussain. "The Utility of Feature Reuse: Transfer Learning in Data-Starved Regimes." *ArXiv*, **abs/2003.04117**, 2020.

[Wal22] Kristin Walter. "What Is Acute Ischemic Stroke?" *JAMA*, **327**(9):885–885, 03 2022.

[WBS04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing*, **13**(4):600–612, 2004.

[WCH20] Zhihao Wang, Jian Chen, and Steven CH Hoi. "Deep learning for image super-resolution: A survey." *IEEE transactions on pattern analysis and machine intelligence*, **43**(10):3365–3387, 2020.

[WGG17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local Neural Networks." 11 2017.

[WHH89] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. "Phoneme recognition using time-delay neural networks." *IEEE transactions on acoustics, speech, and signal processing*, **37**(3):328–339, 1989.

[WHM18] Stefan Winzeck, Arsany Hakim, Richard McKinley, José A. A. D. S. R. Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, Arlindo Oliveira, Youngwon Choi, Myunghee Cho Paik, Yongchan Kwon, Hanbyul Lee, Beom Joon Kim, Joong-Ho Won, Mobarakol Islam, Hongliang Ren, David Robben, Paul Suetens, Enhao Gong, Yilin Niu, Junshen Xu, John M. Pauly, Christian Lucas, Mattias P. Heinrich, Luis C. Rivera, Laura S. Castillo, Laura A. Daza, Andrew L. Beers, Pablo Arbelaezs, Oskar Maier, Ken Chang, James M. Brown, Jayashree Kalpathy-Cramer, Greg Zaharchuk, Roland Wiest, and Mauricio Reyes. "ISLES 2016 and 2017-Benchmarking Ischemic Stroke Lesion Outcome Prediction Based on Multispectral MRI." *Frontiers in Neurology*, **9**, 9 2018.

[WKW16] Karl Weiss, Taghi M Khoshgoftaar, and Ding Ding Wang. "A survey of transfer learning." *Journal of Big Data*, **3**, 2016.

[XBK15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

[XLG21] Fan Xu, Yingying Liang, Wei Guo, Zhiping Liang, Liqi Li, Yuchao Xiong, Guoxi Ye, and Xuwen Zeng. "Diagnostic performance of diffusion-weighted imaging for differentiating malignant from benign intraductal papillary mucinous neoplasms of the pancreas: a systematic review and meta-analysis." *Frontiers in Oncology*, **11**:637681, 2021.

[XSZ21] Kai Xuan, Liping Si, Lichi Zhang, Zhong Xue, Yining Jiao, Weiwu Yao, Dinggang Shen, Dijia Wu, and Qian Wang. "Reducing magnetic resonance image spacing by learning without ground-truth." *Pattern Recognition*, **120**:108103, 2021.

[XZC21] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. "SimMIM: A Simple Framework for Masked Image Modeling." 11 2021.

[YHH20] Jiancheng Yang, Yi He, Xiaoyang Huang, Jingwei Xu, Xiaodan Ye, Guangyu Tao, and Bingbing Ni. "AlignShift: bridging the gap of imaging thickness in 3D anisotropic volumes." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pp. 562–572. Springer, 2020.

[You50] William J Youden. "Index for rating diagnostic tests." *Cancer*, **3**(1):32–35, 1950.

[YYZ] Xiaoxu Yang, Pengxin Yu, Haoyue Zhang, Rongguo Zhang, Yuehong Liu, Haoyuan Li, Penghui Sun, Xin Liu, Yu Wu, Xiuqin Jia, et al. "Deep Learning Algorithm Enables Cerebral Venous Thrombosis Detection With Routine Brain Magnetic Resonance Imaging." *Stroke*.

[YZK22] Pengxin Yu, Haoyue Zhang, Han Kang, Wen Tang, Corey W Arnold, and Rongguo Zhang. "RPLHR-CT Dataset and Transformer Baseline for Volumetric Super-Resolution from CT Scans." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pp. 344–353. Springer, 2022.

[ZCL18] Osama O. Zaidat, Alicia C. Castonguay, Italo Linfante, Rishi Gupta, Coleman O. Martin, William E. Holloway, Nils Mueller-Kronast, Joey D. English, Guilherme Dabus, Tim W. Malisch, Franklin A. Marden, Hormozd Bozorgchami, Andrew Xavier, Ansaar T. Rai, Michael T. Froehler, Aamir Badruddin, Thanh N. Nguyen, M. Asif Taqi, Michael G. Abraham, Albert J. Yoo, Vallabh Janardhan, Hashem Shaltoni, Roberta Novakovic, Alex Abou-Chebl, Peng R. Chen, Gavin W. Britz, Chung-Huan J. Sun, Vibhav Bansal, Ritesh Kaushal, Ashish Nanda, and Raul G. Nogueira. "First Pass Effect." *Stroke*, **49**:660–666, 3 2018.

[ZDP20] Can Zhao, Blake E Dewey, Dzung L Pham, Peter A Calabresi, Daniel S Reich, and Jerry L Prince. "SMORE: a self-supervised anti-aliasing and super-resolution algorithm for MRI using deep learning." *IEEE transactions on medical imaging*, **40**(3):805–817, 2020.

[ZGD21] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises." *Proceedings of the IEEE*, **109**(5):820–838, 2021.

[ZJZ21] Haichen Zhu, Liang Jiang, Hong Zhang, Limin Luo, Yang Chen, and Yuchen Chen. "An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging." *NeuroImage: Clinical*, **31**:102744, 2021.

[ZPN21a] Haoyue Zhang, Jennifer Polson, Kambiz Nael, Noriko Salamon, Bryan Yoo, William Speier, and Corey Arnold. "A Machine Learning Approach to Predict Acute Ischemic Stroke Thrombectomy Reperfusion using Discriminative MR Image Features." pp. 1–4. IEEE, 7 2021.

[ZPN21b] Haoyue Zhang, Jennifer S Polson, Kambiz Nael, Noriko Salamon, Bryan Yoo, Suzie El-Saden, Fabien Scalzo, William Speier, and Corey W. Arnold. "Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time." *Computerized Medical Imaging and Graphics*, **90**:101926, 6 2021.

[ZPY23] Haoyue Zhang, Jennifer S. Polson, Eric J. Yang, Kambiz Nael, William Speier, and Corey W. Arnold. "Predicting Thrombectomy Recanalization from CT Imaging Using Deep Learning Models." 2 2023.

[ZZ19] Zongwei and others Zhou. "Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis." pp. 384–393. Springer International Publishing, 2019.

[ZZZ20] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. "Cross-scale internal graph neural network for image super-resolution." *Advances in neural information processing systems*, **33**:3499–3509, 2020.