Technische Universiteit Delft

# MSc Thesis

## Towards Automatic Cerebral 3D-2D CTA-DSA Registration

Charles Downs

TU Delft

# MSc Thesis

## Towards Automatic Cerebral 3D-2D CTA-DSA Registration

by

Charles Downs

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday February 14, 2025 at 15:45PM.

| | | |
|---|---|---|
| Student number: | 4793862 | |
| Project duration: | January 1, 2024 – February 14, 2025 | |
| Thesis committee: | Prof. dr. ir. M. Reinders, | TU Delft, defense committee chair |
| | Dr. X. Zhang, | TU Delft, supervisor |
| | Dr. T. Höllt, | TU Delft, external committee member |
| | Dr. T. van Walsum | Erasmus MC, daily supervisor |
| | Dr. R. Su | TU Eindhoven, daily supervisor |



TU Delft

# Preface

Throughout this manuscript I describe and discuss my work performed on automated 3D-2D computed tomography angiography (CTA) to digitally subtracted angiography (DSA) registration as part of my masters thesis. This work was conducted in collaboration with the Biomedical Imaging Group Rotterdam (BIGR), at Erasmus MC and the TU Delft Computer Vision Lab.

I would like to thank dr. Ruisheng Su, dr. Theo van Walsum, and Matthijs van der Sluijs for their supervision. Your involvement and enthusiasm for the project was a major driving force for me throughout this thesis. I am particularly thankful to Ruisheng and Theo for the informal and supportive manner of supervision. Ruisheng, I will fondly remember our coffee-break-meetings at the TU Delft library. I would additionally like to express my gratitude to dr. Xucong Zhang, prof. dr. Marcel Reinders, and dr. Thomas Höllt, for being part of the defense committee. Lastly, a heartfelt thank you to dr. Ben Wolf for for the feedback and proof-reading the final version.

I would also like thank Vivek Gopalakrishnan, PhD candidate at MIT, for his collaboration in this work. DiffDRR, the differentiable rendering library developed by Vivek, was a critical component of this whole thesis. Without his work, this thesis would not have been possible. Moreover, I am grateful for the collaboration, when we reached out to suggest improvements to the library (improvements that were critical for our work), Vivek enthusiastically answered and implemented improvements based on our own internal methods, making it possible to use DiffDRR with our own data.

Lastly, I'd like to thank my friends and family for the support. A particular thank you to the members of my band, who provided excellent distraction at times. Our regular practice sessions were one of the few 'stable' weekly obligations I had, and gave me a sense of order in times where things were quite chaotic.

**About the Cover**

Isolating venous structures from the CTA was an important part of the network pre-training for this work. The cover contains an overlay of the arterial phase from a digitally subtracted angiography frame overlaid with the isolated veins from the computed tomography angiography. The images sit roughly in their aligned position, and post-processing was done for an artistic touch.

*Charles Downs*
*Delft, February 2025*

# Contents

# 1 Background

Ischemic stroke is one of the leading causes of death and disability worldwide, particularly in high-income countries, where it ranks second only to ischemic heart disease in terms of disability-adjusted life years [17]. Ischemic stroke occurs when blood circulation is reduced or interrupted due to the occlusion of arteries supplying the proximal anterior circulation [15]. In Western countries, approximately 80% of strokes result from arterial occlusions [5]. The annual incidence of ischemic stroke is substantial: in the Netherlands, 38,000 patients are admitted per year[1], while in the United States, this number reaches 800,000, and nearly 1 million cases occur annually in the European Union [15]. Although stroke-related mortality has declined over the past decade [2], ischemic stroke remains the fifth leading cause of death [2], the primary cause of permanent disability [2], and one of the most common contributors to dementia [2].

## 1.1. Introduction & Motivation

Studies have shown that patient outcome has greatly improved with the inception of endovascular thrombectomy (EVT) [15]. EVT is a medical intervention used in the treatment of ischemic stroke that mechanically extracts the thrombus, which is responsible for the vascular occlusion. The procedure begins by gaining vascular access, typically through a large artery in the groin (such as the femoral artery), and involves navigating a catheter—a long, thin, flexible tube—through the vascular system to the site of the thrombus, as illustrated in Figure 1.1. Catheter navigation is guided by fluoroscopy, typically digitally subtracted angiography (DSA), a technique that provides real-time X-ray images. Once the catheter reaches the site of the occlusion, a device such as a stent retriever is used to engage and remove the thrombus. These devices either ensnare or attach to the thrombus and are carefully retracted, along with the thrombus, through the artery. Sometimes, suction devices are also employed to assist in this process. The main aim of the procedure is to quickly restore blood flow to the affected brain area, thus minimizing brain damage. The effectiveness of the intervention is often measured by the extent of reperfusion, or the restoration of blood flow achieved, as visible in Figure 1.2. After the thrombectomy, patients are closely monitored to manage potential complications, including blood pressure management, prevention of bleeding at the catheter insertion site, and observation for signs of reperfusion injury or recurrent stroke. EVT has shown significant efficacy, especially in patients with large vessel occlusions, leading to improved functional outcomes and reduced disability when performed within a specific time window, typically 6 to 24 hours from symptom onset, depending on individual factors and brain imaging results. This procedure marks a substantial advancement in the management of acute ischemic stroke, providing an option for better recovery in situations previously limited to more conservative treatments like intravenous thrombolysis.

EVT has gained significance since 2015, where 5 clinical trials [3, 10, 16, 4, 13] demonstrated improved patient outcomes after EVT. As a result, since 2015, EVT has been accepted as a de-facto standard care for ischemic stroke patients with large vessel occlusion [9]. These trials collectively identify 3 contributing success factors for EVT: 1) procedure is done with newer generation devices (mainly stent retrievers), 2) more stringent image selection criteria to include only patients with large

[1] Nederlandse Hartstichting, https://www.hartstichting.nl/hart-en-vaatziekten/cijfers-hart-en-vaatziekten

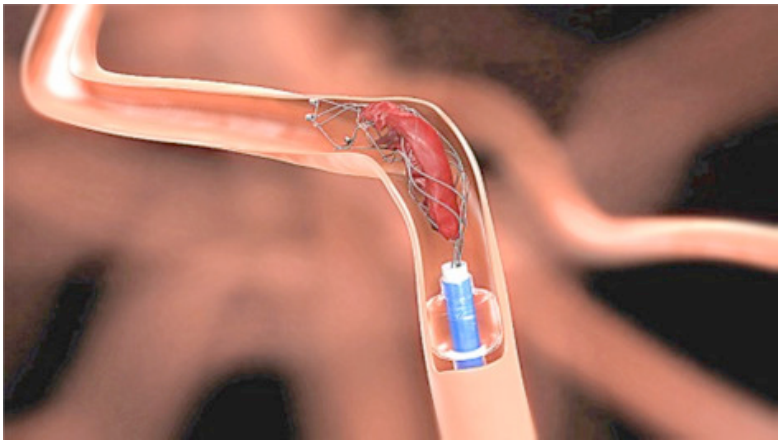


Figure 1.1: Stent retriever, used during EVT to extract the thrombus. Copyright Elsevier

vessel occlusions, and 3) efficient workflows. If these 3 criteria are met, disability rates after acute ischemic stroke, when caused by proximal occlusion of large vessels in the anterior circulation, are significantly reduced [9].

At present, the whole procedure (from device selection to efficient workflow) has been studied and optimized, where typically, stent retrievers are used as an extraction device [15]. In terms of imaging criteria, typically a multimodal computed tomography (CT) or multimodal magnetic resonance imaging (MRI) is performed [15, 17], where the major advantage of CT over MRI is the availability of CT at the emergency room. Patient eligibility for EVT, as well as arterial imaging of cerebral circulation, can be performed, typically with computed tomography angiography (CTA) [17]. DSA imaging then guides the intervention, though are typically limited by the need for contrast injection for vessel opacification.

Finally, Goyal et al. [9] identified *efficient workflows* as one of the key success factors in patient outcome. Such an example could include 3D-2D registration of cross-modality images to aid with the intervention by providing a depth aspect and better global and local information on surrounding structures, as well as occlusion location. 3D-2D cross-modality registration is a type of registration where the images being registered are from different modalities and are of different dimensions. The section below discusses the objective of this paper, where we assess a registration method for cerebral 3D-2D computed tomography angiography to digitally subtracted angiography with applications to EVT.

## 1.2. Purpose of this Thesis

This paper aims to develop and assess a registration method tailored for 3D-2D CTA to DSA registration within the context of EVT. In most cases, a 3D CTA scan is performed to assess patient eligibility for EVT, while 2D DSA is used to guide the procedure. Given the routine availability of both imaging modalities, their registration represents a logical step with the potential to enhance clinical workflows. Exploring such a method aligns with the *efficient workflow* success criterion for EVT identified by Goyal et al [9]. First, it enables the projection of 3D information from CTA onto DSA images, assisting in real-time navigation during interventions. For example, this could include visualizing collateral vessels that are not visible on standard DSA but become apparent in late-phase CTA, potentially indicating the route or trajectory leading to the occlusion. Additionally, it can support the visualization of perfusion-related information such as CTP-penumbra—regions of the brain that are under-perfused, but potentially salvageable during stroke. Tools like a 3D roadmap, generated from registered images, could provide more precise guidance by overlaying critical anatomical landmarks onto the DSA. Such advancements can be valuable in complex cases where vascular structures are partially occluded or unclear in DSA images.

Artificial intelligence (AI) has demonstrated significant potential in image processing, including medical imaging. In particular, convolutional neural networks (CNNs) have proven to be robust for tasks such as classification and segmentation. A growing body of research has explored the application of AI to medical image registration [11, 6]. These methods, however, often require large amounts of labeled training data to achieve high performance, if in a supervised learning setting. The challenges of data

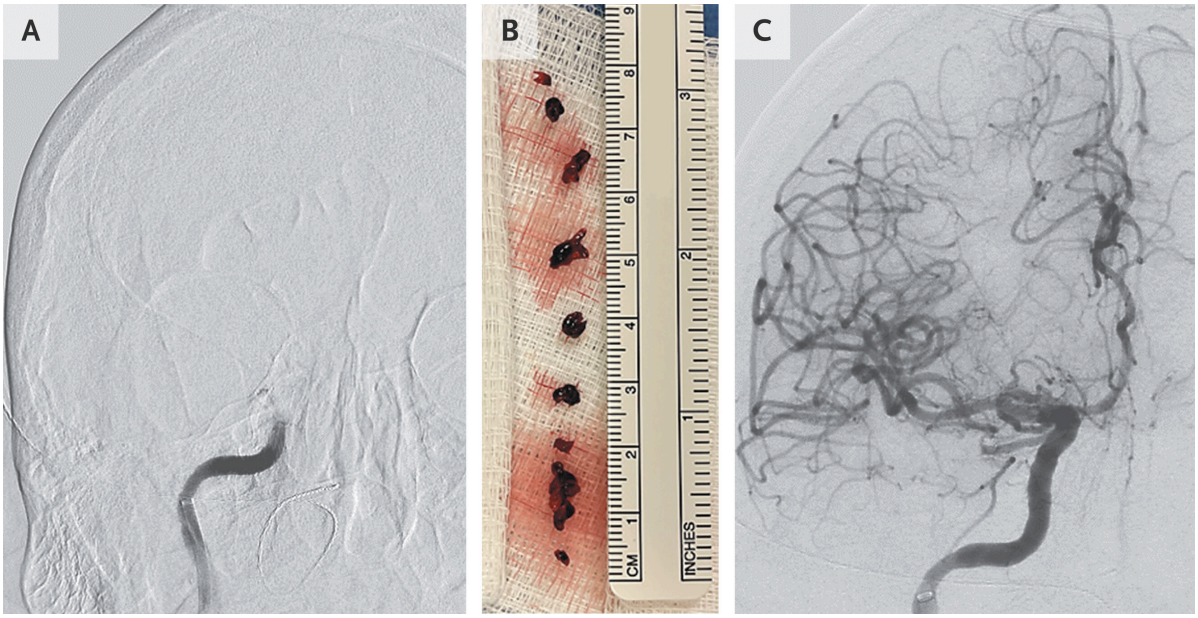



Figure 1.2: Images from the EVT procedure are shown. Panel A displays DSA images capturing contrast injection into the right carotid artery, where blood flow abruptly ceases, indicating the presence of a clot obstructing circulation. Panel B depicts thrombotic material extracted from the occlusion site using a suction device for aspiration. Panel C illustrates a follow-up DSA image after contrast injection at the same location, demonstrating successful restoration of blood flow following clot removal. Reproduced with permission from [14], Copyright Massachusetts Medical Society.

availability in medical imaging are well-documented. In the Netherlands, for instance, patients treated for EVT must provide explicit consent for their data to be used in research. This process introduces logistical hurdles and limits the size of available datasets. Furthermore, once consent is obtained, supervised methods require manual annotations in the form of transformation matrices to register the 3D CTA scans to the 2D DSA images to serve as reference standards. This process demands both medical expertise and a non-negligible time investment. These factors make collecting training data for fully deep learning-based methods time-consuming and resource-intensive. Traditional optimization-based methods offer an alternative that circumvents the need for large datasets. These approaches rely on well-defined metrics to iteratively refine the alignment between images and have been shown to produce accurate results [7, 8, 12]. However, traditional methods suffer from a limited capture range, making them less robust to large initial misalignment. This limitation motivates the exploration of hybrid approaches that combine the strengths of deep learning and optimization-based methods.

In this work, we attempt to address the problem of 3D-2D registration by integrating a CNN with classical optimization techniques. The CNN is used to predict an initial transformation matrix that roughly aligns the CTA to the DSA, leveraging its ability to generalize from limited training data. This initial alignment is subsequently refined through iterative optimization. A hybrid approach is appealing for two main reasons: 1) it reduces the reliance on large training datasets, which are challenging to acquire in medical imaging, and 2) it extends the capture range of traditional optimization methods by starting with a learned initialization. By combining a data-driven initial pose prediction with classical optimization-based methods, such as in [7, 8, 12], the proposed method attempts to produce accurate registrations from limited training data.

We term our method *DeepIterReg*, inspired by its combination of deep learning and traditional optimization. The motivation, methods, implementation, experiments and results are outlined in the research paper provided in Section 2. Supporting and supplementary material to complement the paper is provided in Section 3

# 2

# Research Paper

# Towards Automatic Cerebral 3D-2D CTA-DSA Registration

Charles Alec Downs
TU Delft & Erasmus MC
me@charlesalec.com

## Abstract

*Stroke remains a leading cause of morbidity and mortality worldwide, despite advances in treatment modalities. Endovascular thrombectomy (EVT), a revolutionary intervention for ischemic stroke, is limited by its reliance on 2D fluoroscopic imaging, which lacks depth and comprehensive vascular detail. We propose a novel AI-driven pipeline for 3D CTA to 2D DSA cross-modality registration, termed DeepIterReg. The proposed pipeline integrates neural network-based initialization with iterative optimization to align pre-intervention and peri-intervention data. Our approach addresses the challenges of cross-modality alignment, particularly in scenarios involving limited shared vascular structures, by leveraging synthetic data, vein-centric anchoring, and differentiable rendering techniques. We assess the efficacy of DeepIterReg through quantitative analysis of capture ranges and registration accuracy. Results indicate that our method is able to accurately register 70% of a testset of 20 patients, and is able to improve capture ranges when performing an initial pose estimation using a convolutional neural network.*

## 1. Introduction

Ischemic stroke is one of the leading causes of death and disability worldwide, particularly in high-income countries, where it ranks second only to ischemic heart disease in terms of disability-adjusted life years [24]. Ischemic stroke occurs when blood circulation is reduced or interrupted due to the occlusion of arteries supplying the proximal anterior circulation [18]. In Western countries, approximately 80% of strokes result from arterial occlusions [7]. The annual incidence of ischemic stroke is substantial: in the Netherlands, 38,000 patients are admitted per year[1], while in the United States, this number reaches 800,000, and nearly 1 million cases occur annually in the European Union [18]. Although stroke-related mortality has declined over the past decade [1], ischemic stroke remains the fifth leading cause of death [1], the primary cause of permanent disability [1], and one of the most common contributors to dementia [1].

### 1.1. Endovascular Thrombectomy

Studies have shown that patient outcomes significantly improved with the introduction of endovascular thrombectomy (EVT). [18]. EVT is a medical intervention used in the treatment of ischemic stroke that mechanically extracts the thrombus, which is responsible for the vascular occlusion. The procedure begins by gaining vascular access, typically through a large artery in the groin (such as the femoral artery), and involves navigating a catheter—a long, thin, flexible tube—through the vascular system to the site of the thrombus. Catheter navigation is guided by fluoroscopy, typically digitally subtracted angiography (DSA), a technique that provides real-time X-ray images. Once the catheter reaches the site of the occlusion, a device such as a stent retriever is used to engage and remove the thrombus. These devices either ensnare or attach to the thrombus and are carefully retracted, along with the thrombus, through the artery. The main aim of the procedure is to quickly restore blood flow to the affected brain area, thus minimizing brain damage. The effectiveness of the intervention is often measured by the extent of reperfusion, or the restoration of blood flow achieved. EVT has shown significant efficacy, especially in patients with large vessel occlusions, leading to improved functional outcomes and reduced disability when performed within a specific time window, typically 6 to 24 hours from symptom onset, depending on individual factors and advanced brain imaging results. This procedure marks a substantial advancement in the management of acute ischemic stroke, providing an option for better recovery in situations previously limited to more conservative treatments like intravenous thrombolysis.

### 1.2. Endovascular Thrombectomy Success Criteria

EVT, has gained significance since 2015, where 5 clinical trials [4, 5, 11, 16, 19] demonstrated improved patient outcomes after EVT. As a result, since 2015, EVT has been accepted as a de-facto standard care for ischemic stroke patients with large vessel occlusions [12]. These trials col-

[1] Nederlandse Hartstichting, https://www.hartstichting.nl/hart-en-vaatziekten/cijfers-hart-en-vaatziekten

Figure 1. Example 2D DSA acquisition of arterial circulation.

lectively identify 3 contributing success factors for EVT: 1) *procedure is done with newer generation devices* (mainly stent retrievers), 2) *stringent image selection criteria to only include patients with large vessel occlusions*, and 3) *efficient workflows*. If these 3 criteria are met, disability rates after acute ischemic stroke, when caused by proximal occlusion of large vessels in the anterior circulation, are significantly reduced [12]. At present, the whole procedure (from device selection to efficient workflow) has been studied and optimized, where typically, stent retrievers are used as an extraction device [18]. In terms of imaging criteria, typically a multimodal computed tomography (CT) or multimodal magnetic resonance imaging (MRI) are performed to assess EVT eligibility [18, 24]. Arterial imaging of cerebral circulation during EVT is performed using digitally subtracted angiography (DSA) [24]. DSA imaging then guides the intervention, though are typically limited by the need for contrast injection for vessel opacification.

Finally, Goyal et al. [12] identified *efficient workflows* as one of the key success factors in patient outcomes with EVT. Such an example could include 3D-2D registration of cross-modality images to aid with the intervention, by providing a depth aspect and better global and local information on surrounding structures, as well as occlusion location. We dedicate the following section to cross-modality registration and motivate its usage within the context of EVT and ischemic stroke treatment.

### 1.3. Cross-Modality Registration

Cross-modality registration involves aligning images from different medical imaging modalities to combine their complementary strengths. In EVT treatment, this has the benefit of enabling the projection of 3D information from the CTA onto the 2D DSA, providing better visualization of anatomical landmarks, collateral vessels, and perfusion-related regions like CTP-penumbra. This may further improve catheter navigation in complex cases, assists in identifying occlusion routes, and provides precise guidance through tools like 3D roadmaps. These advancements potentially address the challenges of fluoroscopy's 2D limitations, supporting efficient workflows.

While cross-modality registration is well-studied in medical imaging [8, 13, 17, 25], there is limited research on its application in 3D CTA to 2D DSA registration specifically for EVT. This paper addresses how 3D-2D registration can be approached in the context of EVT. We propose and assess a multi-stage registration pipeline that combines a convolutional neural network (CNN) with traditional optimization-based methods. The CNN first predicts an initial pose for the CTA such that we reduce the initial distance to the registered position. Residual misalignment is then further reduced using an iterative optimization-based approach, as these methods are shown to produce accurate registrations [9, 10, 15]. Our method further leverages DiffDRR [10], a differentiable rendering framework that allows for the analytical computation of gradients during the optimization process. The availability of gradients has shown [10] to improve the optimization speed when when compared to non-differentiable methods, a critical aspect in clinical workflows.

## 2. Related Work

We begin this section with an introduction on how 3D-2D registration is performed in the context of iterative optimization-based techniques—one of the most broadly adopted methods for 3D-2D registration. As inputs, a 3D CTA is given, which we desire to align with a 2D angiogram (in our case, a DSA). The objective is to determine the pose of a CTA relative to a C-arm configuration, maximizing the similarity between its projection and a reference 2D image. The pose of the CTA is parametrized by a position and orientation, each represented by a set of 3 real numbers: $\{r_x, r_y, r_z, t_x, t_y, t_z\}$, with $r_x, r_y$ and $r_z$ the rotation components, and $t_x, t_y$ and $t_z$ translation components. Typically, these parameters are represented as a (rigid) transformation matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \quad , \tag{1}$$

$\mathbf{R}$ corresponds to a rotation matrix according to $r_x, r_y$ and $r_z$, and $\mathbf{t}$ is a vector containing $t_x, t_y$ and $t_z$. A projection matrix is then used to project the resulting CTA according to the pose, resulting in a 2D digitally reconstructed radiograph (DRR). Before discussing these parameters, it is important to outline the geometry of a C-arm system.

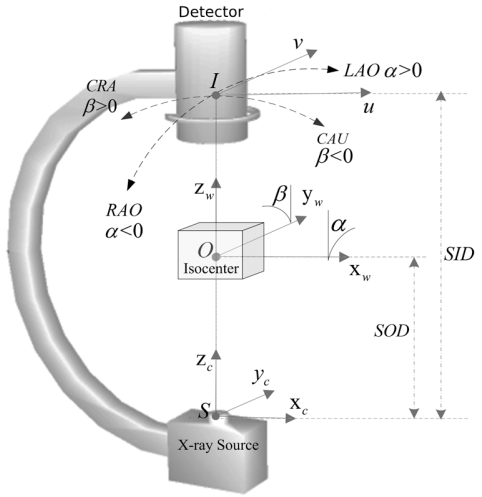



Figure 2. Setup of a (older) typical C-arm system [26].

### 2.1. C-Arm System Geometry

A typical angiography machine, depicted in Figure 2, operates by rotating around two (or three) principal axes, characterized by a primary angle, $\alpha$, and a secondary angle, $\beta$. In addition, the system is defined by intrinsic parameters, such as the source-to-patient (SOD) and source-to-detector (SID) distances, which govern image magnification, resolution, and the position of the isocenter. The isocenter is the center of rotation of the C-arm (which typically is not the midpoint of the source-to-image distance). Three coordinate systems can be discriminated in a C-arm system: the world coordinate system $(x_w, y_w, z_w)$, which we choose as centered at the isocenter, $O$, of the C-arm system. It defines the overall spatial reference, where the rotations $\alpha$ (for Left Anterior Oblique (LAO) and Right Anterior Oblique (RAO)) and $\beta$ (for Cranial (CRA) and Caudal (CAU)) are applied, describing the orientation of the imaging components relative to the isocenter. In DICOM terms, $\alpha$ is the *primary angle*, and $\beta$ the *secondary angle*. The source coordinate system $(x_c, y_c, z_c)$ is centered at the X-ray source $S$, where the X-ray beam originates—this can be seen as the 'camera' coordinate system. The distance from the source to the isocenter is described as the source-to-patient distance. The image plane coordinate system $(u, v)$ is centered at the image detector plane $I$, where the X-ray beam is projected after passing through the isocenter. The axes $u$ and $v$ lie within the detector plane, and the source-to-detector defines the distance between the source $S$ and the detector $I$ in Figure 2.

A projection matrix, $\mathbf{T}$, is used to project the resulting CTA according to its pose and device parameters. Projection matrices in the context of a C-arm system are parametrized by both intrinsic and extrinsic parameters. The extrinsic parameters define the position and orientation of the C-arm relative to the world coordinate system (or 'camera'), and are represented by the transformation matrix $\mathbf{T}$, as shown in Equation 1. The intrinsic parameters characterize the properties of the imaging system, such as the focal lengths, principal point, and source-to-detector distance.

The geometry of the C-arm system directly influences the construction of the projection matrix $\mathbf{P}$, which maps a 3D point $\mathbf{P}_w = (X, Y, Z, 1)^\top$ in the world coordinate system to a 2D point $\mathbf{p} = (u, v)^\top$ on the image plane:

$$\mathbf{p} = \mathbf{P} \cdot \mathbf{K} \cdot \mathbf{T} \cdot \mathbf{P}_w \quad . \tag{2}$$

Here, $\mathbf{K}$ is the intrinsic camera matrix, defined as:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $f_x$ and $f_y$ are the focal lengths along the $u$ and $v$ axes of the image plane, and $c_x$, $c_y$ are the coordinates of the principal point (the center of the image plane).

To compute the projected 2D coordinates $(u, v)$, the homogeneous coordinates $\mathbf{p}_h = \mathbf{P} \cdot \mathbf{K} \cdot \mathbf{T} \cdot \mathbf{P}_w$ are converted to non-homogeneous form:

$$u = \frac{p_x}{p_w}, \quad v = \frac{p_y}{p_w}, \tag{4}$$

where $(p_x, p_y, p_w)$ are the components of the projected point in homogeneous coordinates.

### 2.2. Optimizing a C-Arm Pose

The goal of 3D-2D registration is to optimize the transformation matrix $\mathbf{T}$ such that the digitally reconstructed radiograph generated from the 3D volume projected according to $\mathbf{T}$ aligns with the target 2D angiogram. The optimization setup optimizes a similarity function $\mathcal{L}(\cdot)$ between the DRR and the reference angiogram (e.g., mutual information or normalized cross-correlation). The choice of similarity metric is crucial and dependent on the image modalities (such as multi-modal, versus mono-modal).

The rigid transformation parameters are initialized using an approximate pose constructed from the C-arm parameters. At each iteration, the current transformation matrix $\mathbf{T}$ is used to project the 3D points of the volume onto the 2D plane using the projection model defined in Equation 2, resulting in a DRR. The rigid transformation parameters $\{r_x, r_y, r_z, t_x, t_y, t_z\}$ are updated iteratively using gradient-based optimization techniques, such as stochastic gradient descent (SGD) or Adam. The gradient of the loss with respect to the parameters is computed as:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{T}} = \frac{\partial \mathcal{L}}{\partial I_{\text{DRR}}} \cdot \frac{\partial I_{\text{DRR}}}{\partial \mathbf{T}} \quad , \tag{5}$$

where the term $\frac{\partial I_{\text{DRR}}}{\partial \mathbf{T}}$ captures the effect of pose changes on the DRR, and $I_{DRR}$ is the resulting DRR given the pose parameters. These gradients are used to update the transformation matrix parameters:

$$\mathbf{T}_{k+1} = \mathbf{T}_k - \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{T}} \quad , \tag{6}$$

where $\eta$ is the step size.

The optimization is constrained to the space of rigid transformations, ensuring that $\mathbf{R}$ remains a valid rotation matrix (i.e., orthogonal with a determinant of 1). To enforce this, $\mathbf{R}$ is often parametrized using Euler angles, quaternions, or the exponential map.

While iterative optimization-based methods can achieve accurate registrations, they typically suffer from very limited capture ranges [23]. If the initial pose is too far from the target pose, the optimization process may converge to a local minimum, and therefore an incorrect transformation matrix. We further discuss the importance of capture ranges in a subsequent section.

Given the above specification, the overall optimization objective for 3D-2D registration can be formulated as such:

$$\max_{\mathbf{T}} \mathcal{L}(I_{\text{DRR}}(\mathbf{T}), I_{\text{ref}}) \quad , \tag{7}$$

where $I_{\text{DRR}}(\mathbf{T})$ represents the DRR generated using the current transformation matrix $\mathbf{T}$, and $I_{\text{ref}}$ the fixed reference image.

Hipwell et al. [15] propose an intensity-based 3D-2D registration algorithm for aligning 3D Magnetic Resonance Angiography (MRA) with 2D Digital Subtraction Angiograms (DSA). This method extends traditional 3D-2D registration approaches by focusing on cerebral vasculature alignment for neurointerventions, such as the treatment of arteriovenous malformations and aneurysms. The transformation between the 3D MRA and 2D DSA is modeled as a rigid transformation matrix, $\mathbf{T}$, with six degrees of freedom.

The authors utilize DRRs to simulate 2D X-ray projections of the 3D MRA volume. DRRs are generated by ray-casting through the 3D volume and integrating voxel intensities along the ray paths to approximate X-ray attenuation. Four methods for DRR generation are explored: simple threshold segmentation, direct integration of speed data from phase-contrast MRA, segmentation-based voxel intensity projection, and vessel probability maps. These methods aim to generate DRRs that closely resemble DSAs to enhance the similarity for alignment.

The algorithm optimizes a similarity metric between the DRR and DSA using iterative gradient descent techniques, refining the rigid transformation matrix $\mathbf{T}$ at each step. The similarity metrics evaluated include normalized cross-correlation, mutual information, and gradient correlation, which leverage high-frequency edge information from both the DRR and DSA. To enhance robustness and reduce computational cost, the authors employ a multi-resolution strategy, where the images are progressively refined from lower to higher resolutions, and use concentric region-of-interest masks to constrain the optimization to the region around the vasculature.

Validation on both phantom and clinical datasets demonstrates sub-millimeter accuracy in reprojection errors, with the most robust results achieved using gradient correlation, gradient difference, and pattern intensity similarity measures. The method achieves success rates (reprojection error below 4 millimeters) of 95% for phantom data and 82% for clinical data when initialized within a realistic capture range.

### 2.3. Differentiable Rendering

A critical step in the optimization process described above is the availability of gradients of the loss with respect to the pose parameters, as given in Equation 5. These can either be computed analytically, or by sampling the loss landscape in the direction of interest. Differentiable rendering has the advantage of allowing for the computation of analytical gradients. A contemporary example is DiffDRR, introduced by Gopalakrishnan et al. [10]. DiffDRR generates DRRs from 3D volumes using device parameters like voxel size and spacing, as described at the start of this section. This method incorporates Siddon's method [20] for fast and realistic rendering by simulating X-ray attenuation paths. The authors of DiffDRR implement Siddon's method as a series of vector operations, making it fully differentiable, and facilitating its integration into neural networks for end-to-end training. The projection process involves simulating X-ray generation and attenuation through the 3D volume, and then differentiating the resulting 2D DRR with respect to the *pose* of the CTA (namely, its position and orientation in space). The authors assessed its efficacy in an analogous framework to the optimization framework described above, where the gradients are computed analytically rather than sampled. The proposed method achieves significant speed improvements over traditional approaches. The gradients computed by their method are very close to those obtained via finite differences (within $0.05 \pm 0.01$), but significantly faster to compute, reducing computation time from 73.3 ms to 35.1 ms.

In a subsequent study, Gopalakrishnan et al. [9] applied DiffDRR to a 3D-2D registration problem, registering pre- and intra-operative CT scans. Their approach involved a two-stage process: first, a ResNet18 model was trained on synthetic data generated from pre-operative CTs using DiffDRR, achieving an initial sub-millimeter success rate (SMSR) of 37%. This initial alignment was then refined through iterative optimization, improving SMSR to 87%. The study underscores the importance of a good initializa-

tion in 3D-2D registration, particularly for intensity-based methods, and demonstrates the efficacy of differentiable rendering.

### 2.4. Proposed Method and Contribution

In this paper, we describe a method for 3D-2D CTA to DSA cross-modality registration, which leverages both iterative optimization-based methods, such as those used in [9, 10, 15], in conjunction with a deep learning network. This multi-stage pipeline attempts to overcome the limited capture range of traditional methods [23], while still making use of its ability to generate accurate final registrations. Using a limited dataset consisting of only CTA scans, we train a neural network in a synthetic fashion in order to achieve an initial CTA pose that is within the capture range of iterative optimization-based methods. To this end, we asses the capture range of iterative methods, and investigate to what degree we overcome these capture ranges using a learning-based initialization approach. Using repeat experiments, we then assess the accuracy improvement of each stage of a two-stage iterative optimization step, which further refines the initial CTA pose at each step.

## 3. Method: DeepIterReg

In this section we propose DeepIterReg, illustrated in Figure 3, a multi-stage registration pipeline for 3D-2D CTA to DSA registration, combining a deep learning network with traditional optimization-based methods, powered by DiffDRR's differentiable rendering engine [10]. We posit that, in a similar approach to [9], we can perform a first-step initial registration, such that we obtain an initial pose $\tilde{\mathbf{T}}$, using a deep neural network pre-trained on synthetic data, followed by an iterative registration approach to then compute a second, more accurate pose $\hat{\mathbf{T}}$, where $\hat{\mathbf{T}} \cdot \tilde{\mathbf{T}} \approx \mathbf{T}_{\texttt{target}}$, and $\mathbf{T}_{\texttt{target}}$ is the target rigid transformation (registration) matrix. The initial pose is computed to overcome the limited capture range of traditional iterative-based registration techniques, where $\tilde{\mathbf{T}}$ is the transformation that brings the CTA to its initial pose, and is *ideally* within the capture range of the iterative method employed for the later-stage registration. Effectively, the optimization and learning attempt to find the 6 pose parameters: $\{r_x, r_y, r_z, t_x, t_y, t_z\}$, which we represent as a $4 \times 4$ rigid transformation matrix in all optimization and learning steps, as defined in Equation 1, in an identical manner to the process described in Section 2.2.

At each stage of the pipeline, the objective is to refine the transformation matrix such that it transforms the CTA from its original pose to its target pose. We discuss below how the transformation matrix is optimized in each step of the pipeline.

### 3.1. Learning-based Initialization Step

The initialization step aims to quickly compute an initial transformation, $\tilde{\mathbf{T}}$, that aligns the CTA to an approximate pose close to the target pose, $\mathbf{T}_{\texttt{target}}$. This initial transformation serves as a starting point for further refinement in subsequent steps. We hypothesize that a convolutional neural network (CNN), trained on larger vessels that 'surround' the brain may be sufficient to provide an initial alignment. Visually, venous structures appear to be the most common structure in both modalities, which are typically larger and take on a semicircular structure that contours the skull. Furthermore, venous structures exhibit less inter-patient variability, making them good candidate anatomical features for a learning-based initial registration. However, such an approach would require separating the veins from the arteries in both CTA and DSA. For the DSA, artery-vein separation segmentations are available thanks to the work of Su et al. [22]. However, artery-vein separation is not available for the CTA. In order to obtain vein isolation for the CTA, a combination of morphology and connected components was sufficient to isolate the larger veins that surround the skull. To overcome the modality difference, we binarize the resulting vein segmentation from the CTA, such that both the CTA and DSA are binarized segmentations, as illustrated in Figure 4. These binarized vein segmentations are provided as input to the network during inference. Prior to predicting any initial pose, we can put the CTA in a position and orientation that matches the C-arm configuration during acquisition. We outline how this is performed below.

#### 3.1.1 Pose Initialization

As a first step, we can put the CTA in a 'canonical' pose—that we will refer to hereunder as *radiological* pose—to get a first initial alignment. To this end, we can extend the projection model defined in Equation 2 as such:

$$\mathbf{p} = \mathbf{P} \cdot \mathbf{K} \cdot \mathbf{T} \cdot \mathbf{R}_{2w} \cdot \mathbf{T}_{\texttt{cent}} \cdot \mathbf{P}_w \quad , \tag{8}$$

where $\mathbf{R}_{2w}$ rotates the CTA to the same world coordinate system as the DSA (i.e., the patient orientation of the CTA matches the patient orientation on the table in world coordinates), and $\mathbf{T}_{\texttt{cent}}$ moves the CTA to the origin of the world coordinate system, which coincides with the isocenter of the C-arm. With the CTA at the center position in world coordinates, we can then use the known C-arm angles to rotate the C-arm according to the primary angle $\alpha$ and secondary angle $\beta$, and use the C-arm intrinsic parameters to construct $\mathbf{K}$ in 3. This ensures the initial position of the CTA matches the C-arm configuration during DSA imaging. The registration task can now be seen as correcting for the patients head position. Further details on the chain of transformations applied to the CTA can be found in the supplementary material.

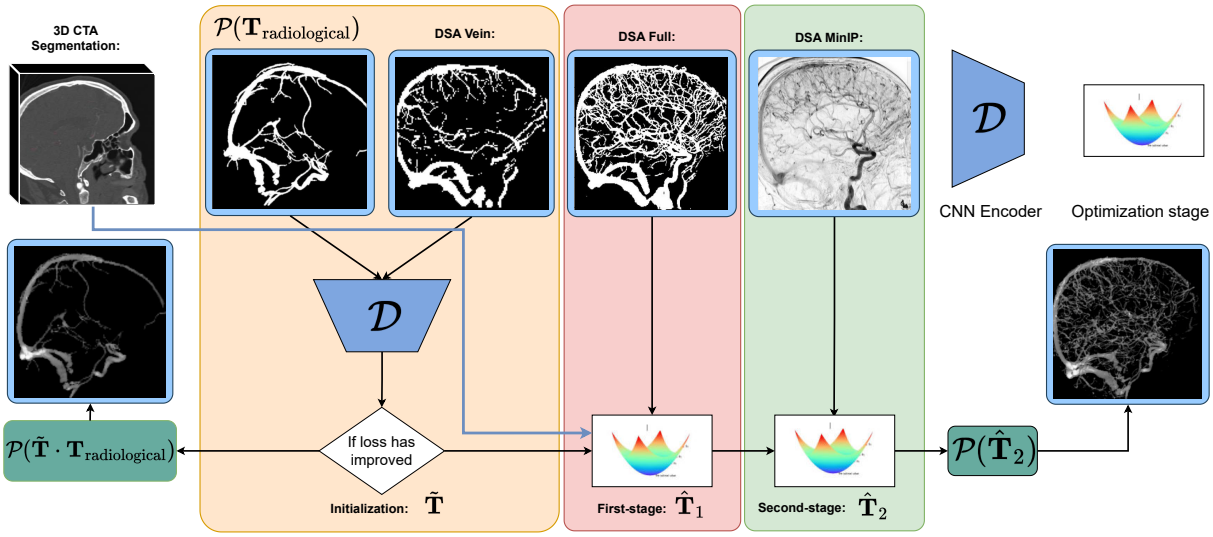
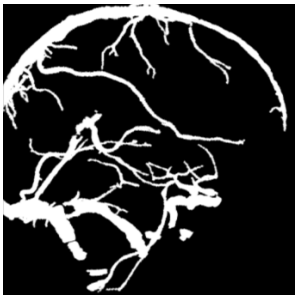



Figure 3. Full pipeline overview. While the 3D CTA is excluded from the diagram, it is included in all optimization stages, the input is therefore a pose, a CTA, and a DSA at both optimization stages.



(a) Example CTA veins obtained by morphology.

(b) DSA vein segmentation from [22].

Figure 4. CTA and DSA veins segmentation comparison. The larger vessels typically exhibit less inter-patient variability and are present in both CTA and DSA, thereby making such structures good candidates for an initial learning-based anchoring.

### 3.1.2 Network Architecture

For the network, we employ a ResNet18 backbone to learn relevant features for the registration task. The 2D projected segmentations, $\text{DRR}_{\texttt{moving}}$ and $\text{DRR}_{\texttt{fixed}}$, are concatenated along the channel dimension, where $\text{DRR}_{\texttt{moving}}$ is the 2D DRR generated from the CTA based on unregistered pose, obtained by applying an offset $\mathbf{T}_{\texttt{offset}}$ to the CTA. $\text{DRR}_{\texttt{fixed}}$ is the fixed DRR in terms of the registration process. Further details on data generation are provided in a subsequent section. The output feature map size from the backbone is $1 \times 512$, which we resize to a 1D feature map, and feed into a final fully connected layer. The output size is 2048, we use $ReLU$ non-linearity activation before regressing the final feature maps via two fully-connected layers, which each predict rotation and translation in a decoupled manner as a pair of 3 floats. A dropout layer is included after $ReLU$ activation. The three fully connected layers have their weights initialized according to He initialization [14] to avoid vanishing gradients from the $ReLU$ activations. The outputs from the last fully-connected layers are then converted into a transformation matrix $\mathbf{T}_{\texttt{pred}}$, which should approximate the inverse of the offset we applied, such that $\mathbf{T}_{\texttt{pred}} \approx \mathbf{T}_{\texttt{offset}}^{-1}$, thereby learning to re-register the CTA the manual offset applied. Figure 5 contains a schematic overview of the architecture. Since we know the range of the perturbations we applied to the CTA, we can further use a $\tanh$ activation on the output layers to normalize them to a $[-1, 1]$ range, and then scale them according to the maximum perturbation magnitude, thereby normalizing the values the network has to learn, in an attempt to improve and stabilize learning. While this stabilizes the training, it will degrade the performance of the network if the transformation is outside the range values the network was trained on. We therefore analyze the limits of the initialization network. We then generate a DRR from the predicted pose using DiffDRR and compute a loss based on the generated DRR and the reference standard DRR from the reference standard, $\mathcal{P}(\mathbf{T}_{\texttt{man}})$.

A schematic overview of the learning process is outlined in Figure 6. The CNN is the ResNet18 backbone we wish to train, who's architecture is given in Figure 5.

We use a weighted combination of Dice loss and two

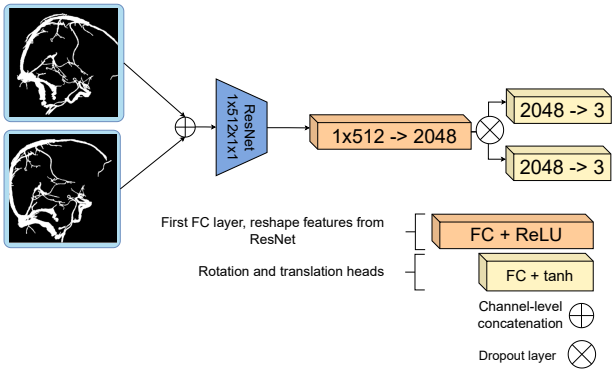



Figure 5. The proposed network, $\mathcal{D}$, extends a ResNet18 architecture to predict a rotation and translation.

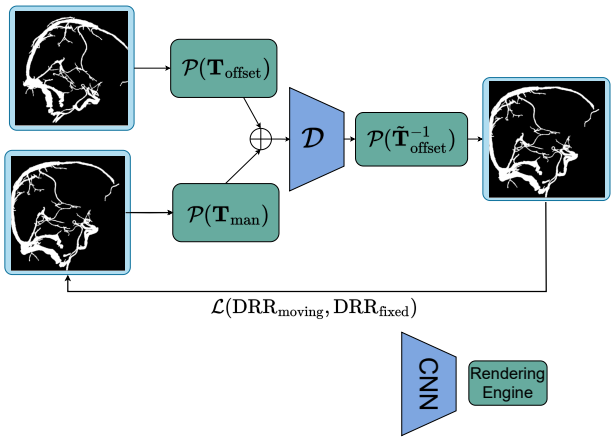



Figure 6. Schematic overview of the learning part of DeepIterReg. $\mathcal{P}$ is an instantiated `drr` class from DiffDRR. DRRs can be generated by supplying a pose as a parameter. The 3D CTA volume is supplied at initialization, and we therefore omit it as a parameter in the above figure.

separate losses on the rotation and translation components, inspired by [9], who attempt to solve a similar 3D-2D registration problem via a deep learning method. The total loss function is given as follows:

$$\begin{aligned}\text{Loss} = &\text{Dice}(\text{DRR}_{\texttt{moving}}, \text{DRR}_{\texttt{fixed}}) \\ &+ \lambda\Big(\mathcal{L}_{\texttt{geo2}}(\mathbf{T}_{\texttt{pred}}, \mathbf{T}_{\texttt{target}}; f) \\ &+ \mathcal{L}_{\texttt{geo}}(\mathbf{T}_{\texttt{pred}}, \mathbf{T}_{\texttt{target}})\Big) \quad , \end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{geo2}}$ is the *double geodesic loss* [6] and $\mathcal{L}_{\text{geo}}$ is the geodesic loss. Both are defined on the special Euclidean group $\mathbf{SE}(3)$, which represents rigid-body transformations consisting of a rotation and translation. These components are detailed below:

$$\mathcal{L}_{\texttt{geo2}}(\mathbf{T}_A, \mathbf{T}_B; f) = \sqrt{d_\theta^2(\mathbf{R}_A, \mathbf{R}_B; f) + d_t^2(\mathbf{t}_A, \mathbf{t}_B)} \quad , \tag{10}$$

$$\mathcal{L}_{\texttt{geo}}(\mathbf{T}_A, \mathbf{T}_B) = \| \log(\mathbf{T}_A^{-1}\mathbf{T}_B)\| \quad . \tag{11}$$

The $\mathcal{L}_{\texttt{geo2}}$ loss separates rotation ($\mathbf{R}$) and translation components ($\mathbf{t}$), allowing for independent penalty scaling. Specifically:

- The translation loss, $d_t(\mathbf{t}_A, \mathbf{t}_B)$, is given by the Euclidean norm of the translation vectors:

$$d_t(\mathbf{t}_A, \mathbf{t}_B) = \|\mathbf{t}_A - \mathbf{t}_B\| \quad .$$

- The rotational loss, $d_\theta(\mathbf{R}_A, \mathbf{R}_B)$, is the geodesic angular distance between two rotation matrices, computed as:

$$d_\theta(\mathbf{R}_A, \mathbf{R}_B) = \| \log(\mathbf{R}_A^T\mathbf{R}_B)\| \quad .$$

  This captures the shortest angular distance on the rotation manifold.

The geodesic loss, $\mathcal{L}_{\texttt{geo}}$, measures the overall misalignment between two transformations $\mathbf{T}_A$ and $\mathbf{T}_B$ on $\mathbf{SE}(3)$. It encapsulates both rotational and translational differences into a single metric by projecting the misalignment onto the tangent space of the manifold (the Lie algebra).

The Dice loss is employed to score the overlap between the predicted DRR and fixed DRR. We opt for a Dice loss as both DRRs generated from the CTAs are given as segmentations, and therefore contain no pixel intensity information.

Finally, in order to perform inference on the trained model, we use real DSA-CTA pairs, the network input therefore becomes $\tilde{\mathbf{T}} = \mathcal{D}(\mathcal{P}(\mathbf{T}_{\texttt{radiological}}), \text{DSA})$, with $\mathbf{T}_{\texttt{radiological}}$ the pose after correcting for the rotation based on the C-arm configuration from the DSA. We refer to this baseline pose as $\mathbf{T}_{\texttt{radiological}}$.

### 3.1.3 Data Generation and Training

Due to the lack of available data, we perform pre-training that relies on simulated perturbations in order to generate the training data. Specifically, we train an encoder network, $\mathcal{D}$, by manually perturbing a registered CTA image with a random transformation $\mathbf{T}_{\texttt{offset}}$ applied from the radiological pose. The network's objective is to predict $\mathbf{T}^{-1}_{\texttt{offset}}$, effectively learning how to invert the applied transformation and re-register the CTA, allowing us to generate an arbitrary amount of data and registrations. As input, we train the network using only the CTA vein segmentations: we first 'synthesize' a *fixed* DSA by projecting the CTA according to its registered pose. A second *moving* CTA is then generated

by projecting the CTA according to its registered pose with the additional offset, $\mathbf{T}_{\texttt{offset}}$. We use DiffDRR to generate the DRRs from the CTA. We call $\mathcal{P}(\cdot)$ the rendering engine from DiffDRR which produces the 2D DRRs, given a CTA and a transformation $\mathbf{T}$, as input, and which is parametrized according to the C-arm configuration[2]. The final DRRs are therefore generated as follows: $\text{DRR}_{\texttt{moving}} = \mathcal{P}(\mathbf{T}_{\texttt{offset}})$, and, $\text{DRR}_{\texttt{fixed}} = \mathcal{P}(\mathbf{T}_{\texttt{man}})$[3].

As the original DSAs are provided as segmentations, we binarize the resulting DRRs to further emulate the downstream real CTA-DSA registration task. An example of the CTA and DSA vein segmentations are given in Figure 4. The venous structured contained in both segmentations are similar and represent similar anatomical features. This motivates training the network on CTA-CTA pairs, as the venous structures are sufficiently similar across both segmentations to potentially generalize to real CTA-DSA pairs, and simplifies the networks task by eliminating non-mutual vessels and overcome the lack of sufficient high-quality DSA images.

### 3.2. Iterative Refinement

Following the initialization of the CTA to an approximate target pose, the subsequent objective is to employ iterative optimization-based methods analogous to what was described in Section 2 such that we achieve an accurate final pose. We can assess whether the predicted pose results in a better alignment by using normalized cross-correlation as a proxy for registration accuracy. A second DRR is generated according to the initial pose, the normalized cross-correlation is then computed between the DSA and the DRR in its radiological pose, and the DSA and the DRR in the predicted initial pose. If the similarity between the DSA and the DRR in its radiological pose is higher, we discard the initial pose as this suggests the initial pose has resulted in a worse registered pose. The radiological pose is then used as an initial pose in the registration process, which is subsequently initiated as outlined in Section 4.3.4. The initial pose is therefore chosen as:

$$\mathbf{T}_{\texttt{init}} = \begin{cases} \mathbf{T}_{\texttt{pred}}, & \text{if } \mathcal{L}(\text{DRR}_{\text{predicted}}, \text{DSA}) \\ & \quad \geq \mathcal{L}(\text{DRR}_{\text{radiological}}, \text{DSA}) \\ \mathbf{T}_{\texttt{radiological}}, & \text{otherwise} \end{cases} \tag{12}$$

[2]We leave the parametrization arguments out of the function parameters for brevity. The rendering engine is initialized according to the C-arm configuration. Generating DRRs via $\mathcal{P}(\cdot)$ therefore also requires the CTA volume: $\mathcal{P}(\mathbf{T}, CTA)$. Further details are given under implementation.

[3]The matrices passed as parameters to the rendering engine are here given as the offsets alone. In the real-world setting, we would need to multiply the offset matrix by the radiological pose in order to apply the offsets to the C-arm configuration: $\mathcal{P}(\mathbf{T}_{\texttt{man}} \cdot \mathbf{T}_{\texttt{radiological}})$.

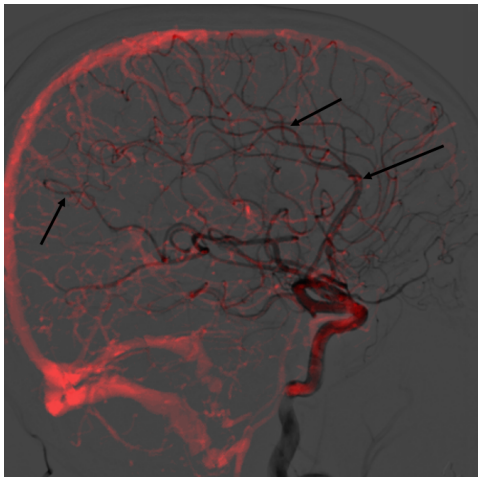


Figure 7. A CTA projected according to its registered pose over a DSA. CTA shown with a red tint to enhance vessel visibility. DSA frame corresponds to arterial phase.

The initialization network, $\mathcal{D}$, having hopefully overcome the limited capture range of such methods. The principal challenge at this stage is to achieve an accurate final registration. As outline previously, we hypothesized that using larger vessel segmentations are sufficient for an approximate initial pose. This is likely not the case for a highly accurate registration—crucially, smaller arteries present in both modalities are typically used when manual registration is performed. This can be illustrated visually, as seen in Figure 7, where the DSA frame is extracted from the arterial phase.

If smaller vessels are critical to an accurate registration, it will be important to make use of the full vessel tree in both modalities. To this end, we propose a two-stage iterative optimization method: an intermediary step after the initialization can be performed to further refine the initial CTA pose. To achieve this, we make use of DiffDRR's differentiable rendering engine in conjunction with the optimization method outlined in Section 2. The optimization algorithm requires as input a fixed reference image, in our case, the DSA, as well as the current pose of the CTA, $\tilde{\mathbf{T}}$, obtained from the initialization. Parameter updates are performed using Adam optimizer until a predetermined number of iterations is reached, and normalized cross-correlation (NCC) as a metric to be optimized. We call the resulting transformation from the optimization stage $\hat{\mathbf{T}}_1$.

Lastly, we posit that a further refinement of the transformation $\hat{\mathbf{T}}_1$ can be achieved by substituting the DSA segmentation with the DSA minimum-intensity projection (MinIP) in the optimization stage defined above, while using an alternative intensity-based loss. The previous optimization approach relied on the segmentation, which effectively discards pixel intensity information from the DSA.

Additionally, the optimization utilized NCC as the loss function, which measures the correlation between the overall structures of the DSA and CTA. However, a potentially more effective refinement could be achieved by substituting NCC with a distribution-based similarity measure, such as mutual information, which accounts for the statistical relationship between the intensity distributions of the two modalities. We therefore define $\hat{\mathbf{T}}_2$ as the second-stage optimization method, which runs in an identical manner to the first-stage, with smaller step sizes in the optimization step, and mutual information (MI) as a similarity metric to optimize. At both steps of the optimization procedure, we select the pose that corresponds to the highest similarity.

# 4. Experiments and Results

In this section, we present the details of our experimental setup, including the data used, the implementation of our registration pipeline, and the evaluation metrics employed. We conduct a series of experiments to assess the performance of our proposed method. First, we outline the data preprocessing steps, including how CTA vein segmentations were obtained, followed by the implementation details of our pipeline. We then evaluate the effectiveness of different optimization strategies and investigate the impact of various vascular anatomies on registration accuracy. To benchmark our approach, we conduct experiments analyzing the capture range of our initialization network and assess the full pipeline's performance across different test scenarios.

## 4.1. Data

We use data from the MR CLEAN registry, a prospective study conducted across 17 centers in the Netherlands, which focused on patients who have undergone EVT for ischemic stroke treatment. The dataset contains CTA scans of each patient, as well as a set of DSA images acquired both pre- and post-EVT. For each patient, an anterior-posterior (AP) scan is available, as well as a lateral scan. A segmentation algorithm was run on the CTAs to produce vein segmentations, similarly, semantic segmentation was performed on the DSA using [22], which produced a set of 3 segmentations for each patient: a full segmentation, a vein segmentation, and an artery segmentation.

A total of 94 lateral-view patient scans were selected from a dataset comprising 182 patients. The dataset includes a combination of pre- and post-EVT DSAs, which were not differentiated for the purposes of this study. The final patient selection was based on the quality of the DSA: suboptimal DSA images, such as ones that suffer from motion, are removed. No selection criteria are applied to the CTA images. Manual annotations (in the form of rigid transformation matrices) are provided by an in-house medical researcher, in collaboration with a team of Clinical Medicine students and a Biomedical Engineering master's student. The registrations are performed in an in-house designed MeVisLab module. The total dataset size used for this paper consisted of 81 training patients, 9 validation patients, and 20 testset patients.

## 4.2. Metrics

In order to assess the accuracy of the resulting registrations, we can examine the average Euclidean distance between the final registered pose, and the reference standard, or Mean Projection Error (MPE). This is accomplished by projecting a hypercube (8 evenly spaced points from the CTA) using both the reference standard registration matrix as well as the registration matrix from the optimization algorithm and measuring the projection error, as defined below.

$$\text{MPE} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{p}_i^{\text{ref}} - \mathbf{p}_i^{\text{reg}}\|_2 \quad , \tag{13}$$

where $\mathbf{p}_i^{\text{ref}}$ are the points projected according to the reference standard, and $\mathbf{p}_i^{\text{reg}}$ the points projected according to the obtained registration matrix. An accurate registration should project the points to approximately the same coordinates, making the overall average distance near zero. For multiple registrations, a scatterplot can be constructed where we plot MPE before registration versus MPE after registration. This metric provides a quantitative assessment of how well the registration algorithm aligns the perturbed CTA to the reference standard in terms of projection error.

For a large number of perturbations, it may be difficult to visually infer the capture range from a scatterplot. Therefore, when evaluating the capture range with a large number of perturbations (such as with the simulated poses) we construct a histogram and plot the median deviation of the points projected *before* versus the points projected *after* registration. We define the bin size such that there are 10 total bins, each containing 10% of the registrations. An optimal registration will result in the bin lying on the $y = 0$ line, meaning that the transformation matrix has moved the points to a distance of 0 to the reference standard. The capture range can be identified by the point where the registrations diverge from the $y = 0$ line. We further supplement the plot with confidence intervals.

## 4.3. Implementation

This section describes the implementation details of our proposed registration pipeline, including preprocessing steps and computational methods used, such as CTA vein extraction and optimization. We detail the process of generating DRRs and provide an overview of the iterative registration process and the integration of the deep learning-based initialization step.

#### 4.3.1 CTA Vein Segmentation

Artery-vein semantic segmentation is not available for the CTA images. As outlined in Section 3.1, we perform an initial registration with the vein segmentation on both the DSA and the CTA. In order to obtain the vein segmentation from the CTA, we apply a morphological opening to remove small, thin structures such as minor arteries or noise by convolving the volume with a spherical structuring element of size 1. Following this, connected component analysis is performed to label distinct regions. To further refine the mask, components smaller than a predefined volume threshold of 1000 voxels are filtered out, ensuring that only larger anatomical structures remain. The resulting binary mask is converted back to the NIFTI file format, preserving the spatial metadata of the original image. This is performed as a pre-processing step so as to not add extra computations during training or optimization.

#### 4.3.2 Iterative Optimization Method

The iterative optimization algorithm is central to the registration pipeline. It improves the alignment by iteratively updating the initial transformation predicted by the CNN. To this end, we employ the Adam optimizer in PyTorch with two distinct learning rates ($\eta$ in Equation 6). The optimization process is embedded in its own function, `optimize`, which takes as arguments the learning rates, an initial pose (in the form of a `Registration` class, outlined below), as well as the specified number of iterations. The `optimize` function is built in an analogous way to the sample code provided in DiffDRR's documentation. We therefore refer the reader to the DiffDRR documentation for further details[4]. The version of DiffDRR used is version v.4.0.0.

#### 4.3.3 Generating DRRs

DRR generation is performed using DiffDRR's `DRR` class, which is initialized according to the C-arm configuration. The C-arm parameters are directly extracted from the DICOM file headers. The arguments to be passed to the `DRR` class are: the CTA file, which is read using TorchIO's `Subject` class, the desired pixel spacing, and the source-to-detector distance. When instantiated, the `DRR` class can be used to generate arbitrary DRRs based on a transformation matrix, $\mathbf{T}$, which transforms the CTA and generates the corresponding DRR.

To generate random synthetic DRRs, which is required during training and for evaluating capture ranges, we obtain a set of 6 floats, $\{r_x, r_y, r_z, t_x, t_y, t_z\}$, corresponding to a rotation ($r$) and translation ($t$) offset. The rotation parameters for the synthetic DRRs are sampled from $\mathcal{U}(-35, 35)$ in degrees, and translation parameters sampled from $\mathcal{U}(-45, 45)$ in pixels. These parameters are then converted to a $4 \times 4$ rigid transformation matrix using DiffDRR's `pose_from_carm` function. To benchmark different registration methods, such as optimization-only versus with an initialization, we save the offset transformation to a comma-separated variable file such that they can be re-used for future experiments.

In order to generate DRRs according to their radiological pose, $\mathbf{T}_{\texttt{radiological}}$, or to their registered pose (if available), $\mathbf{T}_{\texttt{man}}$, the transformation matrices can be used to generate the corresponding DRRs by passing the pose as an argument to the instantiated `DRR` class: $\texttt{drr}(\mathbf{T}_{\texttt{radiological}})$. The `DRR` class is effectively the rendering engine, which is initialized according to the C-arm, and then corresponding DRRs can be generated by passing transformation matrices as arguments. In Figure 6, the instantiated `DRR` class is referred to as $\mathcal{P}(\cdot)$.

#### 4.3.4 Registration Process

The registration method combines the optimization method, outlined in Section 4.3.2 and DRR generation, outlined in Section 4.3.3. DiffDRR conveniently provides a `Registration` class which is used intermediately between the registration and optimization. The `Registration` class is provided an instance of the `DRR` class and an initial pose for the CTA. The class instance is called at each iteration of the optimization and generates a DRR according to the current pose parameters. The instantiated `DRR` class is then passed as an argument to the `optimize` function, as well as the objective function to be maximized.

The registration process consists of two passes, the `optimize` function is therefore used twice with different parameters. The first-pass registration is performed using the DSA segmentation and NCC as a similarity metric. For this first-pass, we set the learning rates: $lr_{\text{rot}} = 10^{-2}$ for rotational components and $lr_{\text{xyz}} = 1$ for translational components. The resulting pose is then further refined by calling the `optimize` function a second time, with mutual information as a loss, and by substituting the DSA segmentation for the DSA MinIP. The learning rates are adjusted to $lr_{\text{rot}} = 10^{-4}$ and $lr_{\text{xyz}} = 0.04$.

#### 4.3.5 Initialization Network and Training

We employ the standard ResNet18 architecture as implemented in the TorchVision library with batch normalization. The standard ResNet18 architecture is designed to have an input channel dimension $C$ of one. The first convolutional layer is therefore adjusted to take an input with a channel dimension $C$ of two for the concatenation of both DRRs.

[4] https://vivekg.dev/DiffDRR/

We use a learning rate of $0.001$ for the optimizer. To prevent overfitting, a dropout layer with $p = 0.1$ is incorporated after the backbone. Synthetic DRRs used for training are generated as specified in Section 4.3.3. The parameters for $\mathbf{T}_{\texttt{offset}}$ are sampled uniformly from $\mathcal{U}(-25, 25)$ pixels for translation and $\mathcal{U}(-15, 15)$ degrees for rotation and are used to generate a moving CTA, $\text{DRR}_{\texttt{moving}}$.

The loss function used for training is defined in Equation 9, with a weighting parameter $\lambda = 0.001$, motivated by the choice of weight used for a similar registration task in [9]. The Dice loss is used from [22] and the Geodesic losses are implemented in DiffDRR's metrics class.

After training, and therefore during inference, the initialization can be achieved by passing a *real* DSA vein segmentation image and a DRR of the CTA in its radiological pose.

### 4.4. Experiment: Anatomy for Registration

The purpose of this experiment is to determine which vascular anatomies yield different registration results. For instance, to establish if there is a trade-off between capture range and accuracy when using the vein segmentation or the full segmentation for the registration process. Using the 20 testset patients, the starting poses of the CTAs are randomly generated within a know distance to the reference standard registration pose. For each registration we save the rotation and translation applied to the CTA in order to be able to apply the same perturbation on the vein-only based registrations, thereby allowing for a direct comparison between the two. Results are analyzed by observing the mean projection error (MPE) for a subset of points to the reference standard CTA.

The results are shown in Figure 8. The median capture range can be understood as the point where the median distance *before* diverges from the $y \approx 0$ line. Empirically, we find that a distance below 5 pixels corresponds to an accurate final registration, suggesting that the median capture range is 20 pixels for both veins and full segmentations. For the full segmentation, the MPE before registration, averaged over **all** patients, is 22 pixels, or 19 mm, whereas after, it is reduced to 14 pixels, or 12.9 mm. For the vein-based registration, we find that the overall MPE after registration is 14 pixels, or 12.7 mm.

### 4.5. Experiment: Optimization Method Capture Ranges

To evaluate the capture ranges of the optimization method, we run a registration of the CTA to the DSA using the full dataset of 94 patients. The final accuracy of the registration is determined using MPE to the reference standard. This experiment compliments the previous experiment by performing registrations of the CTA from their radiological pose, as opposed to sampled registrations where the starting pose is randomly generated. While performed on a limited size dataset, this experiment potentially provides for a more realistic distribution of poses, and therefore of capture range.

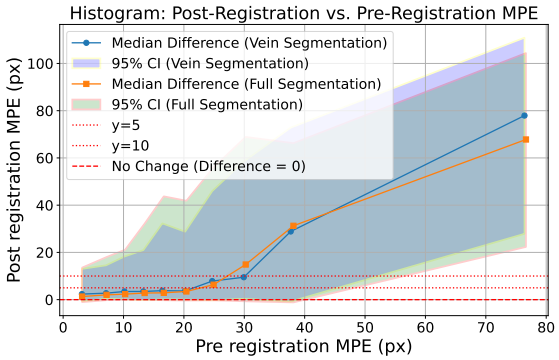



Figure 8. Comparison of simulated registrations, using the full segmentation versus using the vein segmentation only

Figure 9 illustrates the average Euclidean distances before and after registration when registering the 94 patients from their radiological pose.

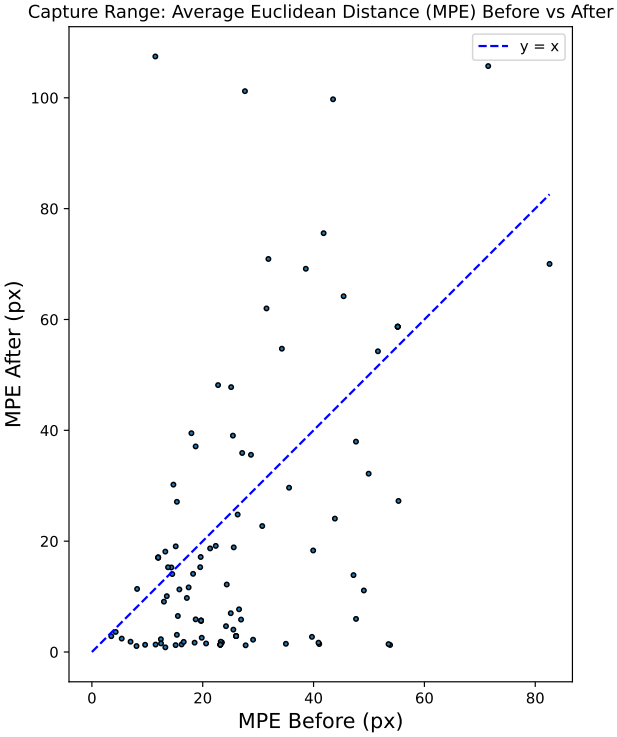



Figure 9. Average Euclidean distance (MPE) for all patients.

We observe an MPE *before* registration of 26 pixels, or 22.7 mm, and an MPE after of 21 pixels, or 18.3 mm.

Figure 10 contains the histogram of median deviations

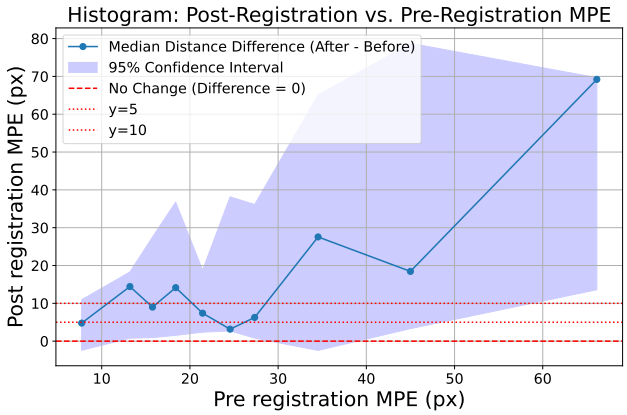
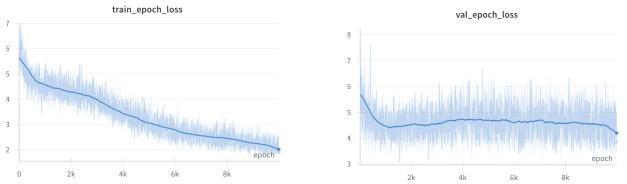



Figure 10. MPE before versus after registration for all 94 patients using optimization-only registration.

corresponding to Figure 9. For smaller distances, such as distances below 25 pixels, the median oscillates between 0 and 15, and eventually diverges at 25 pixels, which approximately corresponds to what was discovered using simulated poses. Over all patients, taking a $y = 5$ MPE threshold, we find that 37 of the 94 patients are successfully registered, corresponding to 39% of of the dataset.

## 4.6. Experiment: Network Training

The encoder network, $\mathcal{D}$, was trained for a total of 10000 epochs on an Nvidia Titan Xp GPU. The training and validation losses are given in Figure 11. The loss curves for each individual loss component are given in the supplementary material. We use a batch size of 1, with 16 batches per epoch. A validation step is then performed after each epoch on the full validation dataset.




(a) Training loss (b) Validation loss

Figure 11. Train and validation losses for the initialization network. Dark blue line corresponds to smoothed loss, as the small batch size results a noisy loss curve.

## 4.7. Experiment: Accuracy of Full Pipeline

In this experiment, the full pipeline performance—and therefore the improvement from the initialization—is assessed. We perform a two-fold assessment. First, using the set of 20 testset patients, we observe whether the number of successfully registered patients improves. Second, we apply perturbations to the 20 testset patients to increase the total number of registrations. This larger number of registrations can then be used to assess the median capture range improvement.

### 4.7.1 Testset Registrations

We first evaluate the results of the 20 testset patients when registered via the optimization-only method, illustrated by the blue points in Figure 13. Each point in the scatter-plot corresponds to a single patient, which is given by the label on each point. Using a success threshold of 5 mm, there are a total of 11 success cases. We find the overall MPE after registration is 1.94 pixels, or 1.71 milliliters for success cases. The overall MPE over all patients is reduced from 24 pixels, or 21.14 millimeters to 11 pixels, or 9.78 millimeters.

The contribution of the initialization network is evaluated by analyzing the reduced distances in Figure 13, which highlight an improvement over the optimization-only approach. The horizontal dotted lines illustrate how the distance to the reference standard changes when the initialization is applied. If the initialization improves the final registrations, each point in the scatterplot should shift downward toward $y \approx 0$. All registrations below the $y = 10$ except for 10153 have an optimal loss, meaning that they have a maximum theoretical loss by the end of the optimization process, thus making the total number of successful registrations 14, or 70%. The overall MPE is also lowered to 8 pixels, or 7.62 millimeters. For successfully registered patients, the overall MPE is 2.44 pixels, or 2.13 millimeters, to the reference standard.

### 4.7.2 Simulated Registrations

We can further supplement these experiments by investigating the net improvement to capture range brought by the initialization in simulated registrations. Using the 20 testset patients, we can make an identical plot to Figure 8 and observe if the point of divergence is extended beyond 20 pixels.

Figure 12 illustrates the change in median capture range resulting from the addition of the initialization step. Compared to the optimization-only approach, we observe that the median capture range starts to diverge from $y = 5$ at 40 pixels, a net improvement over the 20 pixel median capture range identified without using an initialization.

## 4.8. Experiment: MinIP-Based Refinement

To assess whether an intensity-based refinement using mutual information as a loss improves the final registration result, we run the three-stage pipeline on the same set of CTA-DSA pairs as used in experiment 4.7.1 and observe if the MPE decreases.

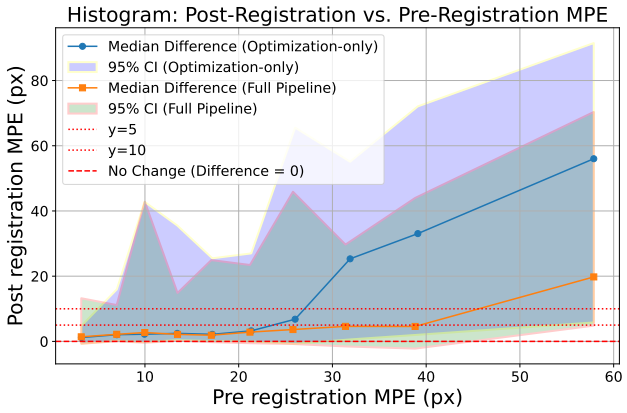



Figure 12. Capture range plot, optimization-only versus full pipeline.

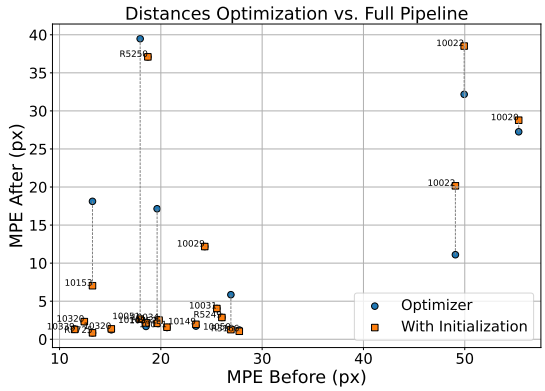



Figure 13. Comparison of Euclidean distances: optimizer-only versus with deep learning initialization. Dotted line shows how distance change.

With the addition of a third-stage intensity-based iterative refinement, we find that the MPE over all successfully registered patients decreases to 2.08 millimeters, while the number of successfully registered patients remains 14.

### 4.9. Experiment: Manual Assessment

Finally, a visual assessment is performed by a clinical expert, this experiment aims to determine if a radiologist can reliably discriminate between registration results. This provides a qualitative assessment of the registration, and also can be used to assess the quality of the manual registrations (our reference standard) compared to the automatic registrations. To this end, we asked a neurointerventional radiologist to score the automatic registrations versus the benchmark manual reference standard registrations.

The radiologist was supplied with a blind 2-way test. Our sample consists of the 14 successful registrations from Section 4.7.1. Each comparison is either between the two-stage and three-stage, the two-stage and the reference standard, or the three-stage and the reference standard. The radiologist was asked to rank each comparison, resulting in a total of $14 \times 3$, or 42 comparisons. The scale used is given below:

Table 1. Preferences given to each method by an interventional radiologist.

| Method | Number of Times Preferred |
|---|---|
| Three-stage | 13 |
| Two-stage | 9 |
| Reference standard | 6 |

- **Scale 1**: left definitely better than right
- **Scale 2**: left is better than right
- **Scale 3**: left is the same as right
- **Scale 4**: right is better than left
- **Scale 5**: right definitely better than left

Table 1 summarizes how many times each method was preferred by the interventional radiologist. A Wilcoxon Signed Rank Test was used to determine whether the radiologist's rankings revealed statistically significant differences between methods. The test identified a significant difference ($p = 0.0146$) between the three-stage versus the reference standard, suggesting that the three-stage registration offers meaningful improvements over the reference standard. More details on the assessment are provided in the supplementary material.

## 5. Discussion

This paper describes DeepIterReg, a multi-stage registration pipeline designed to align 3D CTA to 2D DSA images with applications to endovascular thrombectomy. The method combines a deep learning-based initialization with an iterative optimization step. The deep learning initialization attempts to overcome the limited capture range of traditional optimization methods, while the second stage optimization processes are used to achieve a final accurate registration. Using a dataset of CTA-DSA pairs, the method was evaluated for accuracy and capture range in both simulated and real scenarios. Experimentally, we find that our method is able to increase the median capture range of the traditional optimization-based method from 20 pixels to 40 pixels, as well as increase the number of successfully registered patients from 11 to 14 in a testset of 20 patients. Combining deep learning for pose initialization with traditional optimization techniques has seen more interest recently since the inception of DiffDRR [10], a differentiable

DRR rendering engine that can be used in conjunction with deep learning methods. In this paper, we described an alternative deep learning setup for use in registration pipelines. We attempt to 'simplify' the registration task for the deep learning step by training the network in a mono-modal manner, and only on venous structures from CTA segmentations, due to being mutually present in both CTA and DSA modalities. Our hypothesis was that an initial anchoring of the CTA to the DSA is possible using only larger vessels (veins), and can be achieved by synthetic pre-training on simulated poses with CTA vein segmentations. This novel training approach circumvents the needs for large amounts of high quality CTA-DSA pairs. DSA images typically suffer from motion artifacts, making them unusable in registration settings and as training data. These hypotheses are supported by the fact that larger venous structures provide sufficient anatomical landmarks to establish an initial alignment, as they are typically present in both modalities and exhibit lower inter-patient variability than smaller arterial structures. By restricting the deep learning model's focus to these shared structures, the problem is effectively simplified, allowing the network to produce a coarse yet meaningful initial transformation that generalizes to real CTA-DSA pairs. We refer the reader to the supplementary material for a visualization of the initial vein-based anchoring produced by the network. Since the network is only tasked with approximating an initial registration, it does not require highly accurate reference standard registrations, thereby reducing the need for time-consuming manual annotations. Additionally, approximate registrations relying solely on venous structures are easier and faster to obtain, as aligning smaller arteries is unnecessary.

In the paper we highlighted the importance of high-quality CTA-DSA pairs. If the quality requirements are not met, the registration results may suffer. The testset contains 3 CTAs with quality issues: two are missing large amounts of the veins due to clipping, while one has an abnormally large rotational offset. Adjusting the dataset to exclude these 3 patients, and only include high quality data, the success rate for our method becomes 77%. This however does suggest that even with only high-quality data, the method is not able to register all testset patients.

While training the network in a synthetic manner circumvents the need for high-quality CTA and DSA images during training, our method relies on intermediary segmentation algorithms: a learning-based method for the DSA, and a morphology-based method for the CTA. These intermediary steps can produce poor quality segmentations, even if the DSA or CTA are of high quality, thereby limiting the performance of the method. In clinical practice, this makes the method very difficult to interpret. The claim cannot be made that the method works with high-quality data due to the possibility of a failure in the intermediate processing. More generally, the method relies on a lot of data pre-processing. Extensive testing would have to be done to identify the limits and constraints of each step before one can make any clinical viability claims.

Separately, while we argue that a small distance to the reference standard is necessary for the optimization step to accurately register the CTA, Figure 9 suggests that in some cases in the $x < 20$ range, even if the distance to the reference standard was small before registration, the registration may still fail. This suggests there is a need for more thorough testing of optimization stage, as well as potentially improving the optimization stage itself, such as including schedulers, which we discuss in the supplementary material. Conversely, two patients with distance above 40 pixels were still accurately registered.

While other methods exist for 3D-2D registration that combine deep learning with iterative optimization [3, 9], it remains a relatively under-researched area, particularly with cross-modality registration, which this paper attempts to address. This indicates that the field of 3D-2D registration, particularly approaches that integrate deep learning with traditional methods, is still in its infancy, highlighting the need for further research and development. Particularly the use of larger datasets, or directly training on real CTA-DSA pairs, as opposed to the synthetic approach adopted in our method and in [9]. Additionally, further research into pre-training based on simplified datasets, or in a synthetic manner, may also offer solutions to initial pose estimation with other modalities and different datasets, as demonstrated in our method.

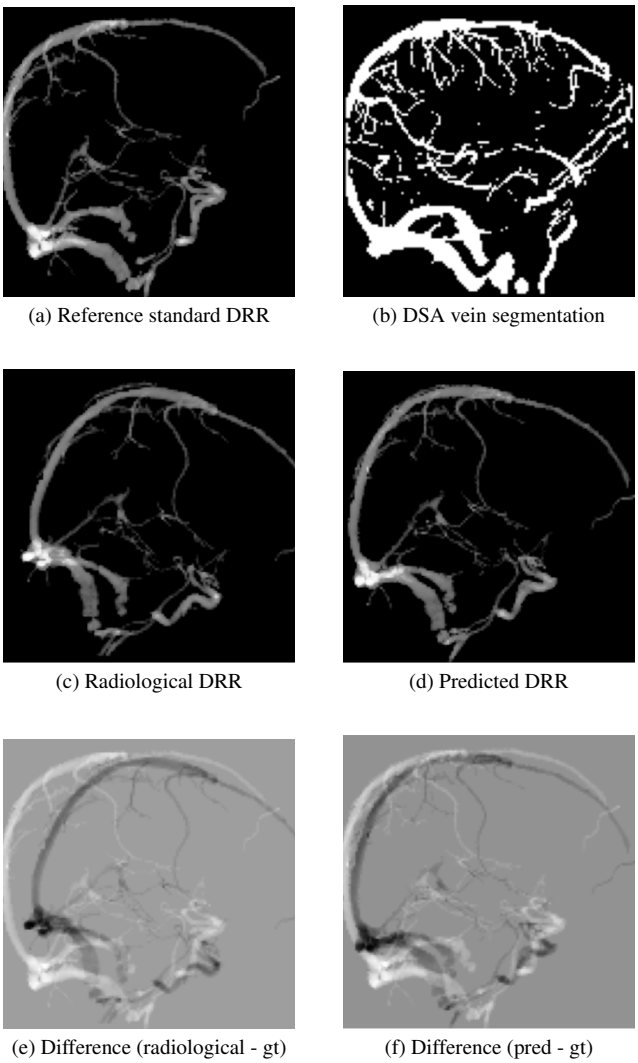
The clinical relevance of this work is also subject to discussion. While 3D-2D registration has potentially valuable applications for EVT, many of the factors stated above hinder its usage in real-world scenarios. In its current state, the method may be usable as a tool for aiding in decision making, assuming the CTA and DSA images obtained for the patient are usable and produce a successful registration. An interventional radiologist can then deem the result as accurate and usable, and hence use it to aid in decision making, or discard it as non-usable. Furthermore, real-world clinical applications would potentially require some level of user interaction. For instance, the size of the structuring element used to segment the CTA veins may need to be tuned to produce usable CTA vein segmentations. It is well documented that the adoption of (AI) decision aiding tools in healthcare practices is slow [2, 21] and that there is still a lack of acceptance and trust in such decision-aiding tools [21]. Providing a tool that requires to be tweaked by a healthcare professional is therefore likely to face difficulties with adoption. Our tool assumes that a CTA is performed in order to assess patient eligibility. While this is the case in the Netherlands, it is not always the standard patient selection imaging modality worldwide [12]. Lastly, the total

runtime for one registration on an Nvidia Titan Xp is approximately 10 minutes. Whether or not this is a viable in an emergency-room setting is yet to be determined. How our method fits into a clinical workflows, and how it can be optimized for such a usage should form part of a future viability study. Further hyperparameter optimization may offer solutions to lowering the number of iterations at each optimization stage.

Concerning the study, the majority of the testset data (16 out of 20, or 80%) consisted of data from a single center where imaging protocols are very stringent. This often results in high quality CTA-DSA pairs, meaning the testing of the method made use of data mostly emanating from a single-center with stringent imaging protocols. Furthermore, 46% of the training data also emanates from the same center, potentially biasing the training and removing a domain-shift introduced by a large varied dataset. As a further consequence, a large portion of the dataset is single-vendor data. Consequently, larger and varied datasets are imperative to further improve the network and assessment of the method. However, as the network only relies on approximate reference standard registrations, there is potential to create larger datasets with a lesser time involvement. Lastly, the network was trained on $128 \times 128$ resolution images, an 8-fold downsampling factor. During inference, subsampling DSA images to such a resolution can harm the quality of the final images. It introduces subsampling artifacts that are not present during training due to the DRR rendering process. Training the network on higher-resolution images, such as $256 \times 256$ or larger, could enhance its ability to generalize to real CTA-DSA data. Future work could focus not only on expanding the dataset but also on utilizing higher-resolution images to investigate whether this improves the accuracy and robustness of the initial pose estimation.

During this study, we attempted to identify which vascular anatomies are important for registration tasks. We investigated whether there is a trade-off between accuracy and capture range when using the full vessel segmentation versus only the vein segmentation. While our results showed no conclusive effect of veins versus the full segmentation in the registration process, we did not investigate the effect of registering with arteries only, or with the removal of random vessels from the segmentation as a pre-processing step. Furthermore, we did not distinguish between using pre- and post-EVT images. Pre-EVT images typically reveal less vascular structure, due to the lack of collateral blood flow. Identifying which vascular anatomies are relevant for registration is crucial, as there is no certitude to obtain a full vessel tree in a clinical setting. Our testset included both pre- and post-EVT cases, and we did not identify any difference in the registration results. Further research could attempt to identify what vascular anatomical structures are conducive to an accurate registration. This may allow to have a classification of patients that *can* be registered, versus patients that cannot.

Lastly, the study may benefit from including multiple graders in the visual assessment of the registrations. We queried one radiologist who had a preference for the automatic registrations; more conclusive results are needed, which would require querying multiple radiologists, as well as checking inter-observer reliability.



(a) Reference standard DRR (b) DSA vein segmentation

(c) Radiological DRR (d) Predicted DRR

(e) Difference (radiological - gt) (f) Difference (pred - gt)

Figure 14. Example inputs, outputs and reference standard DRRs for the initialization network, $\mathcal{D}$.

## 6. Conclusion

This paper introduces DeepIterReg. A multi-stage registration pipeline that leverages deep learning for an initial pose estimation, followed by iterative optimization to refine the registration. We find that including a deep learning

network improves the median capture range from 20 to 40 pixels, and increases the number of successfully registered patents from 11 to 14, or 70%, out of a testset of 20 patients. Correcting for low quality CTA segmentations in the testset, the method is able to successfully register 77% of the testset patients, suggesting there is a need for high-quality CTA and DSA segmentations for reliable and accurate results.

## References


[1] AHA. Heart disease and stroke statistics-2017 update: A report from the american heart association. *Circulation*, 135:e146–e603, 2017. 1

[2] Molla Imaduddin Ahmed, Brendan Spooner, John Isherwood, Mark Lane, Emma Orrock, and Ashley Dennison. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*, 15(10), 2023. 14

[3] Gabriel De Araujo, Shanlin Sun, and Xiaohui Xie. Adaptive image registration: A hybrid approach integrating deep learning and optimization functions for enhanced precision, 2024. 14

[4] Olvert A Berkhemer, Puck SS Fransen, Debbie Beumer, Lucie A Van Den Berg, Hester F Lingsma, Albert J Yoo, Wouter J Schonewille, Jan Albert Vos, Paul J Nederkoorn, Marieke JH Wermer, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *New England Journal of Medicine*, 372(1):11–20, 2015. 1

[5] Bruce CV 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, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *New England Journal of Medicine*, 372(11):1009–1018, 2015. 1

[6] Gregory S Chirikjian. Partial bi-invariance of se (3) metrics. *Journal of Computing and Information Science in Engineering*, 15(1):011008, 2015. 7

[7] Valery L Feigin, Carlene MM Lawes, Derrick A Bennett, and Craig S Anderson. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. *The lancet neurology*, 2(1):43–53, 2003. 1

[8] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, 2020. 2

[9] Vivek Gopalakrishnan, Neel Dey, and Polina Golland. Intraoperative 2d/3d image registration via differentiable x-ray rendering, 2023. 2, 4, 5, 7, 11, 14

[10] Vivek Gopalakrishnan and Polina Golland. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Clinical Image-based Procedures: 11th International Workshop, CLIP 2022, Held in Conjunction with MICCAI 2022, Singapore, Proceedings*, Lecture Notes in Computer Science. Springer, 2022. 2, 4, 5, 13

[11] 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, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *New England Journal of Medicine*, 372(11):1019–1030, 2015. 1

[12] 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. 1, 2, 14

[13] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31:1–18, 2020. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[15] John H Hipwell, Graeme P Penney, Robert A McLaughlin, Kawal Rhode, Paul Summers, Tim C Cox, James V Byrne, J Alison Noble, and David J Hawkes. Intensity-based 2-d-3-d registration of cerebral angiograms. *IEEE transactions on medical imaging*, 22(11):1417–1426, 2003. 2, 4, 5

[16] 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, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *New England Journal of Medicine*, 372(24):2296–2306, 2015. 1

[17] Francisco PM Oliveira and Joao Manuel RS Tavares. Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014. 2

[18] Panagiotis Papanagiotou and George Ntaios. Endovascular thrombectomy in acute ischemic stroke. *Circulation: Cardiovascular Interventions*, 11(1):e005362, 2018. 1, 2

[19] Jeffrey L Saver, Mayank Goyal, Alain Bonafe, Hans-Christoph Diener, Elad I Levy, Vitor M Pereira, Gregory W Albers, Christophe Cognard, David J Cohen, Werner Hacke, et al. Stent-retriever thrombectomy after intravenous t-pa vs. t-pa alone in stroke. *New England Journal of Medicine*, 372(24):2285–2295, 2015. 1

[20] Robert L Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical physics*, 12(2):252–255, 1985. 4

[21] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care, 2023. 14

[22] Ruisheng Su, P. Matthijs van der Sluijs, Yuan Chen, Sandra Cornelissen, Ruben van den Broek, Wim H. van Zwam, Aad van der Lugt, Wiro J. Niessen, Danny Ruijters, and Theo van Walsum. Cave: Cerebral artery–vein segmentation in digital subtraction angiography. *Computerized Medical Imaging and Graphics*, 115:102392, 2024. 5, 6, 9, 11

[23] Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The

impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. *Frontiers in Robotics and AI*, 8:716007, 2021. 4, 5

[24] H Bart Van der Worp and Jan van Gijn. Acute ischemic stroke. *New England Journal of Medicine*, 357(6):572–579, 2007. 1, 2

[25] Max A Viergever, JB Antoine Maintz, Stefan Klein, Keelin Murphy, Marius Staring, and Josien PW Pluim. A survey of medical image registration–under review, 2016. 2

[26] Xuehu Wang, Jian Yang, Yang Chen, Danni Ai, Yining Hu, and Yongtian Wang. Optimal viewing angle determination for multiple vessel segments in coronary angiographic image. *IEEE Transactions on Nuclear Science*, 61(3):1290–1303, 2014. 3

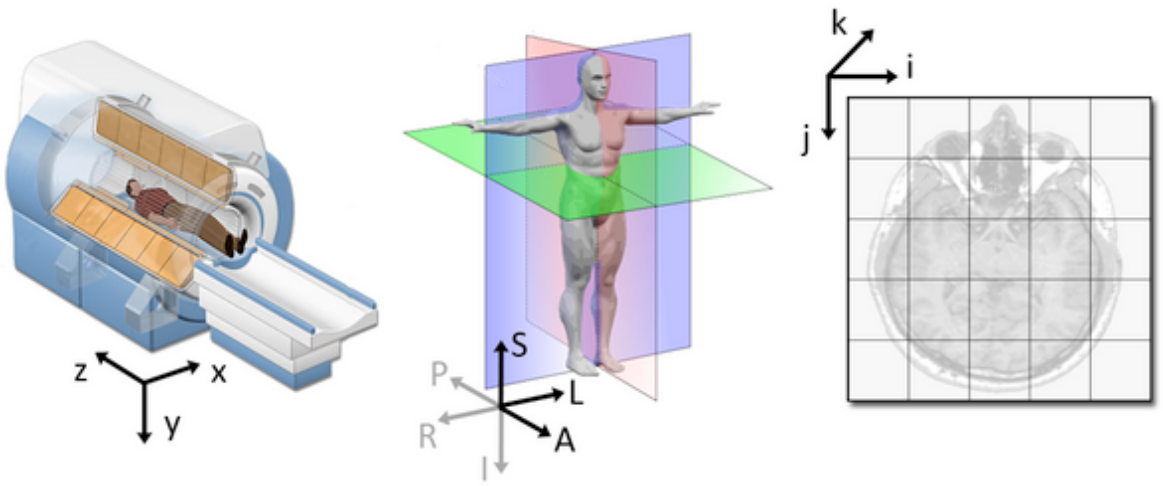
# 3

# Supplementary Material

## 3.1. Pose Distribution and Statistics

For all experiments that involved simulated registrations, random pose parameters were sampled from a uniform distribution, such that $\{r_x, r_y, r_z, t_x, t_y, t_z\} \in \mathcal{U}(-r, r)$ in pixels or degrees. As a result, each individual rotation and translation component was assigned a random float within this range, these sampled values effectively correspond to an offset applied to the patients head. The position of the patient can be better understood by observing the coordinate systems typically present in CT imaging systems, as provided in Figure 3.1.

Figure 3.1: Coordinate systems of a CT system [1]




As briefly mentioned in the research paper, it would not be unreasonable to assume that there is more variation in rotation parameters around the $L$-axis from Figure 3.1 rather than the $S$- or $A$-axis. The reasoning being that one would expect that a patient is more likely to have a 'slouching' head, where the head is tilted left or right. This can be better analyzed by deconstructing the manual transformations matrices and observing the magnitude of each individual rotation and translation component. This is illustrated in Figure 3.2. Looking at the 95th percentile groups, it is clear that certain components, namely the $X$ and $Y$ components for rotation, and the $X$ component for translation, have larger values relative to the CT-system world coordinates. This confirms that for rotation, the patients head is more likely to be looking left or right, or tilted.

This suggests that poses sampled during training could be be more realistic if sampled according to the distribution of poses reflected by the values found by deconstructing the transformation matrices, as opposed to sampling in a random uniform manner, thereby providing for a more realistic training setup.

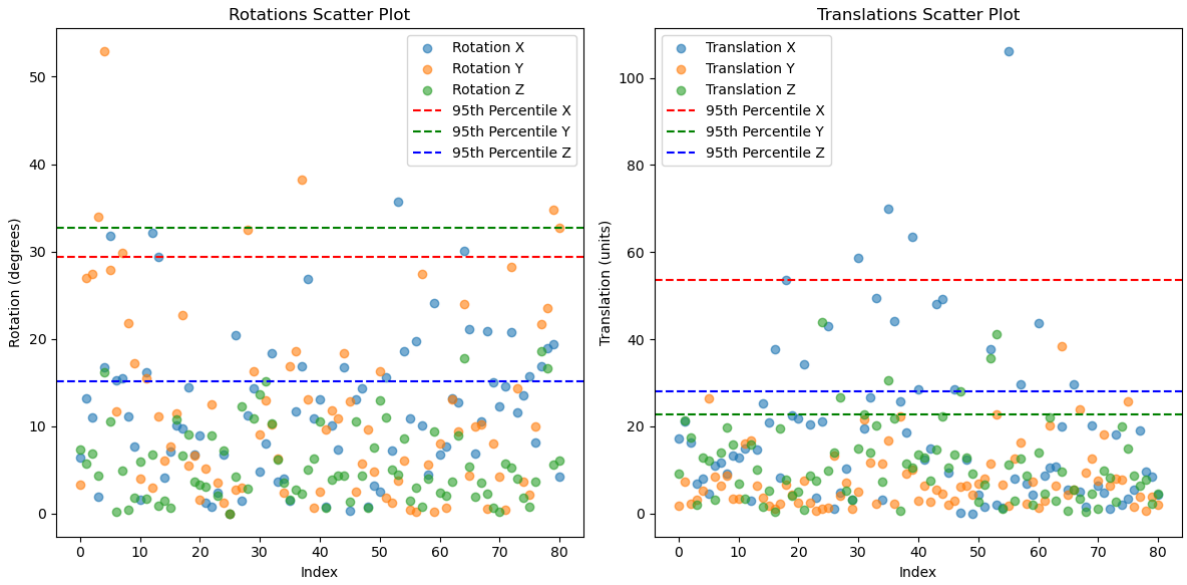



Figure 3.2: Scatterplot of individual rotation and translation components from the manual registration matrix.

## 3.2. Data

### 3.2.1. Training Data

The total dataset size used for this paper consisted of 81 training patients, 9 validation patients, and 20 testset patients. For a patient to be a candidate for the proposed method, high-quality CTA and DSA pairs are needed. Moreover, the pipeline makes use of the vessel segmentation from the DSA, as well as the vein segmentation for the CTA. It is frequently observed that the DSA images suffer from motion artifacts, resulting in sub-par segmentations, which in turn harms the performance of the method.

In terms of deep learning, 81 training patients can be considered a relatively small sample size, particularly given the complexity of the task of predicting a pose. Access to larger amounts of training data has the potential to improve the performance of the initialization network, which may be desirable, particularly as the confidence intervals for the registration plot are wide.

### 3.2.2. Data Quality

While improving the initialization network is desirable, it does not solve the overarching issue that high-quality data is imperative for the method to work adequately. While we propose a system that has overall satisfactory performance, its usage is heavily limited by patient motion during DSA acquisition. Unfortunately, there is no direct solution to this. Patients undergoing EVT are typically aware and conscious, limiting patient motion while a patient is having a stroke has no straightforward solution. This suggests our method has the possibility of being used in some cases, but is far from general clinical viability.

## 3.3. Training Losses

Our choice of loss is guided by similar work on 2D-3D registration using deep learning in the paper proposed by [7]. In this work, the authors propose the following loss:

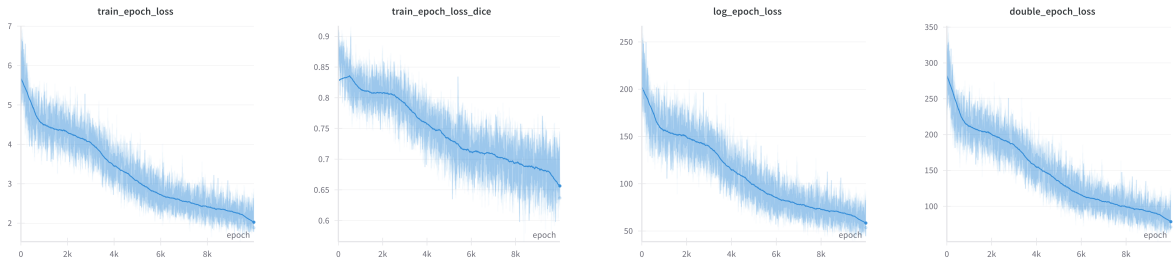
$$-\mathcal{L}_{\text{mNCC}}(\mathbf{I}, \hat{\mathbf{I}}) + \lambda_1 \mathcal{L}_{\log}(\mathbf{T}, \hat{\mathbf{T}}) + \lambda_2 \mathcal{L}_{\text{geo}}(\mathbf{T}, \hat{\mathbf{T}}) \quad , \tag{3.1}$$

where $\mathcal{L}_{geo}$ is the *double geodesic* loss on $\mathbf{SE}(3)$, and $\mathcal{L}_{log}$ the geodesic loss, outlined below:

$$\mathcal{L}_{\text{geo}}(\mathbf{T}_A, \mathbf{T}_B; f) = \sqrt{d_\theta^2(\mathbf{R}_A, \mathbf{R}_B; f) + d_t^2(\mathbf{t}_A, \mathbf{t}_B)} \quad , \tag{3.2}$$

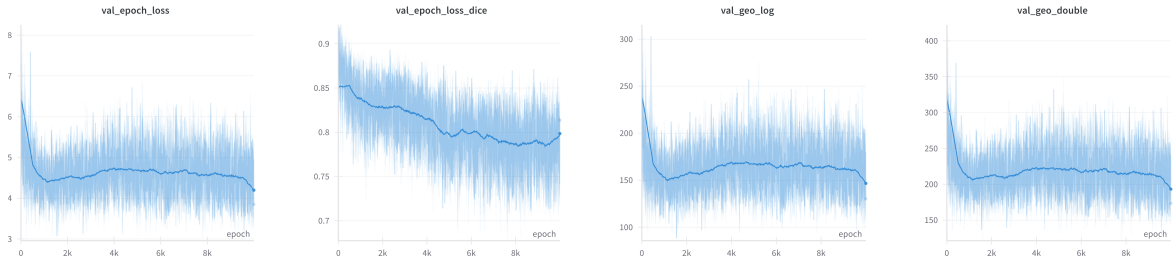
$$\mathcal{L}_{\log}(\mathbf{T}_A, \mathbf{T}_B) = \| \log(\mathbf{T}_A^{-1} \mathbf{T}_B) \| \quad . \tag{3.3}$$

Above, $d_t$ is the norm between the two translation components, and $d_\theta$ the angular distance between the axis of rotation. $\mathbf{T}$ corresponds to the pose matrix, while $\mathbf{R}$ is the rotation component from $\mathbf{T}$, and $\mathbf{t}$ the translation component. This loss combines an image similarity metric, in this case normalized cross-correlation, with two additional losses on the transformation matrices. Due to the similar nature of the problem setup in [7], we adopt a similar loss, where NCC is substituted for a Dice loss, as the proposed training setup uses segmentations, and $\lambda_1 = \lambda_2 = 0.01$. During training, we logged the loss at each iteration for both training and validation steps, where the figures below contain the loss at each iteration. The loss curves are given in Figure 3.3 and validation loss in Figure 3.4.




(a) Overall loss (b) Dice loss (c) Log Geodesic loss (d) Double Geodesic Loss

Figure 3.3: Loss curves during training




(a) Overall loss (b) Dice loss (c) Log Geodesic loss (d) Double Geodesic Loss

Figure 3.4: Validation loss curves during training

The log geodesic, as well as double geodesic display a rapid decrease in the early epochs of the training. This suggests that dynamically adjusting the $\lambda$ weight parameters may be conducive to stabilizing the learning, potentially improving training.

In their paper, [7] adopt normalized cross-correlation (NCC) as their similarity metric in the loss function, where the objective is to align pre- and intra-operative X-rays, meaning both images are from the same modality, thereby making NCC a suitable similarity measure. Due to the similarity between [7] and our own method, we initially also opted for NCC as a similarity metric in the loss. However, when training with a non-segmented CTA and its segmented DSA counterpart, we observed no reduction in loss throughout the training process. Upon replacing NCC with the Dice similarity coefficient and applying thresholding to the CTA, we achieved improved training loss, which ultimately enabled the network to train successfully in its current configuration. Based on this experience, we do not recommend using NCC during training, as it failed to facilitate effective learning in our setup.

## 3.4. Sample Registrations

Due to the relatively small size of the testset—totaling 20 patients—it is possible to investigate the registration process for a subset of the patients. We can additionally plot the loss from each of the 20 patients, as given in Figure 3.5. In the majority of cases, the first stage has converged in under 500 iterations. However, there are cases where the loss still improves after 1500 iterations. For this reason, when testing the method on real data, we choose 2000 iterations.

The loss behavior exhibited is far from smooth—there are large oscillations in short timeframes. Experiments using learning rate schedulers, such as reducing the learning on plateau, and reducing

on plateau with warm restarts, did not stabilize the optimization. We frequently observe that while a loss may plateau for a very long time, it may still have a rapid increase in later iterations, which we illustrate in a sample registration below. Using a scheduler has the undesirable result that it would be difficult for the optimization to escape the local minima that corresponds to the plateau. Including learning rate restarts in conjunction with reducing the learning rate on plateau did solve these issues in some cases, but ultimately when tested on the full 20 testsets, did not result in an overall improvement in the number of success cases. Overall, we suspect that there is room for further optimization of the learning rates as well as the potential to include a scheduler to stabilize the optimization. A hyperparameter search, while extremely time consuming, may ultimately benefit and improve the optimization. A hyperparameter search was not performed in this case due to the time required for such a search. Assuming a grid search with the objective of finding the best pair of learning rates, $\eta_1$ for rotation and $\eta_2$ for translation, if we were to evaluate 5 values for $\eta_1$ and 5 values for $\eta_2$, we would need to perform 25 parameter searches. In a more general case, if we have $N$ different hyperparameters, and we wish to evaluate $k$ different values for each hyperparameter, the total number of parameter searches required in a grid search would be $k^N$, which quickly becomes infeasible as $k$ becomes larger. The optimization process over a single patient for 2000 iterations has a total runtime of approximately 10 minutes when running on an Nvidia Titan XP. This total execution time has to then be multiplied by the number of patients, *and* by the number of parameters in the search space. This quickly becomes extremely time-consuming, particularly if there are hardware limitation constraints.

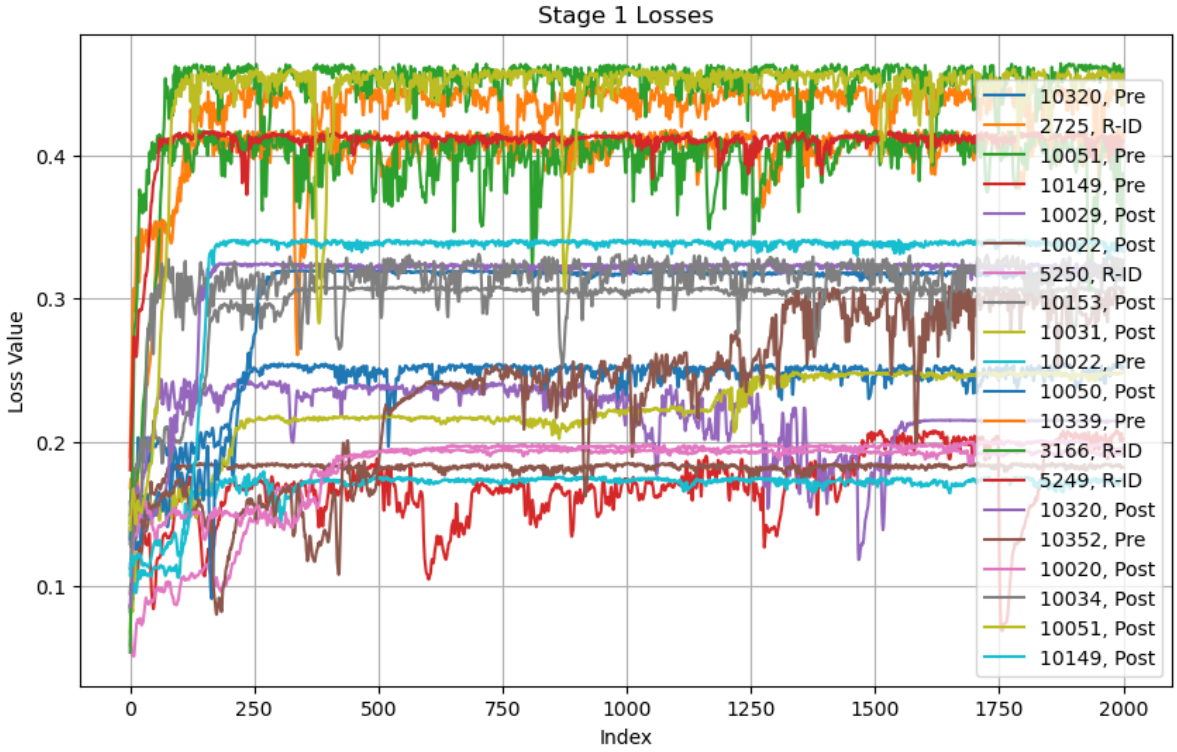



Figure 3.5: Losses from first stage NCC-based registration

The losses for all patients during the second stage are given in Figure 3.6. The change in losses are subtle and much less stable in the long-term than in the first stage. In most cases, there is an initial rapid increase in similarity in the first iterations. Subsequently, some patients have a slowly—yet noisy—increase in loss. Others have a flatlined loss throughout the full optimization procedure, suggesting the result from the first stage is already optimal. These loss trajectories also suggest further hyperparameter optimization can be performed on the second stage. Moreover, in many cases, the rapid loss increase occurs in the first 25 iterations (as illustrated in individual cases below). This suggests that in many cases, allowing the optimization to run for the full 2000 iterations is not needed. In order to reduce the total runtime, it would be desirable to detect and exit the optimization process at an early stage if the loss has flatlined.

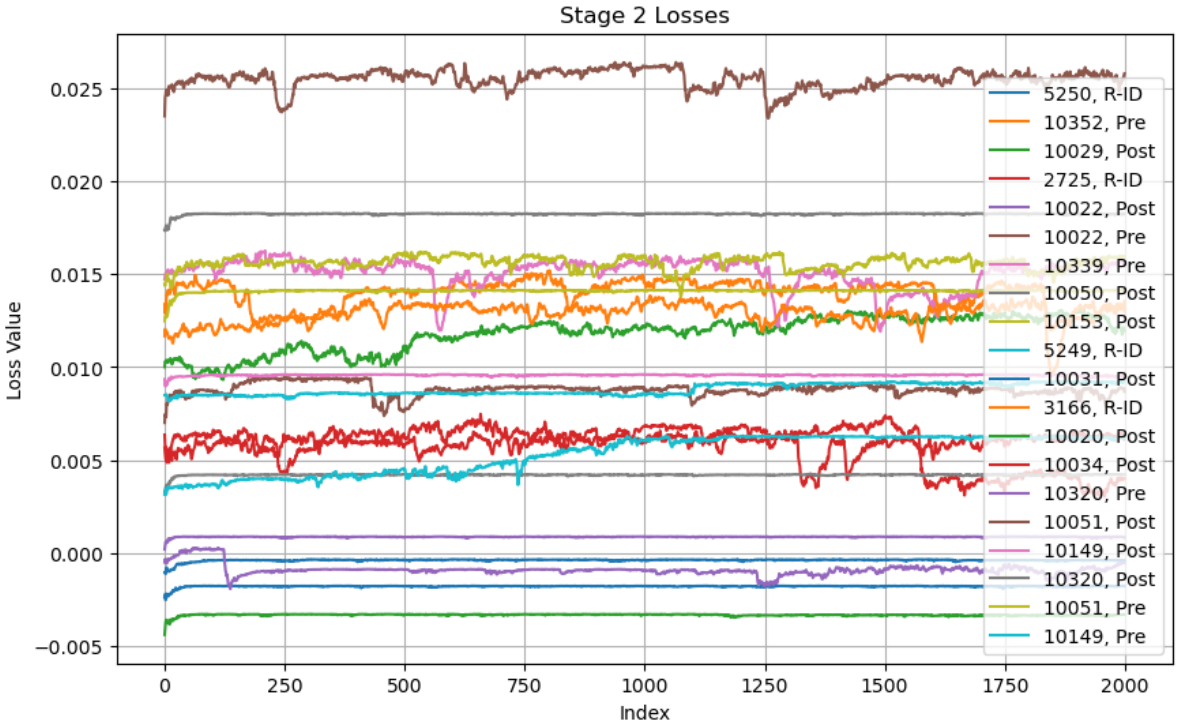



Figure 3.6: Losses from the mutual-information based second stage optimization.

### 3.4.1. Registration Example: Patient 10034

Figure 3.7 contains DRRs generated from registering patient 10034 from its radiological pose. The input to the network, as well as inputs to the optimization stages are provided. Additionally, the two sets of 8 points used to compute the Mean Projection Error (MPE) are plotted in the resulting DRR, and their distances are illustrated in green. The associated losses from the first and second stages are given in Figure 3.8.

The reference standard transformation matrix, $\mathbf{T}_{\text{ref}}$, and the matrix obtained from the optimization process, $\mathbf{T}_{\text{pred}}$ are given in Equation 3.4. All values in $\mathbf{T}_{\text{ref}}$ except for the $z$-component of the translation vector, $t_z$, are very close to their reference standard values. We observe in most cases that the largest difference in all components is $t_z$, this is most likely due to the difficulty of optimizing the 'depth' of the pose. We find that manually adjusting $t_z$ has a minimal impact on the similarity score, making it difficult to optimize for. While the $t_z$ difference can seem large, it becomes negligible when computing the MPE, as illustrated in Figure 3.7.

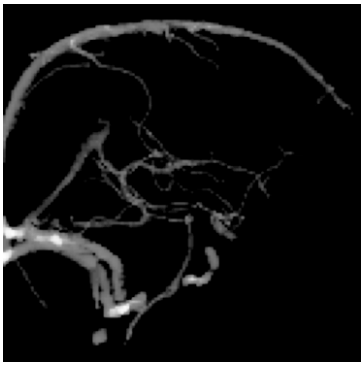
$$\mathbf{T}_{\text{ref}} \qquad\qquad \mathbf{T}_{\text{pred}}$$

$$\begin{bmatrix} -0.0279 & 0.9983 & 0.0517 & -732.5620 \\ -0.9597 & -0.0412 & 0.2779 & 24.4803 \\ 0.2796 & -0.0418 & 0.9592 & 28.5226 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} -0.0377 & 0.9956 & 0.0859 & -728.4005 \\ -0.9657 & -0.0584 & 0.2531 & 38.0633 \\ 0.2570 & -0.0735 & 0.9636 & 50.6748 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.4}$$

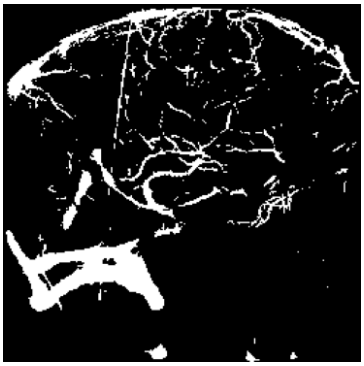
### 3.4.2. Registration Example: Patient 10352

The DRRs and DSA images used are provided in Figure 3.9, and respective losses in Figure 3.10.

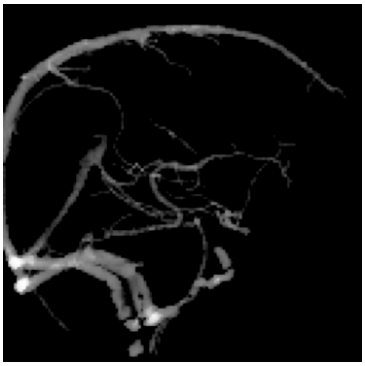
The loss from the second stage for patient 10352 (PRE) in Figure 3.10 does not seem to provide any meaningful improvement. Given the small distance to the reference standard, it is possible that the output from the first stage already provides the highest achievable accuracy from any of the pipeline stages.



(a) Moving CTA.

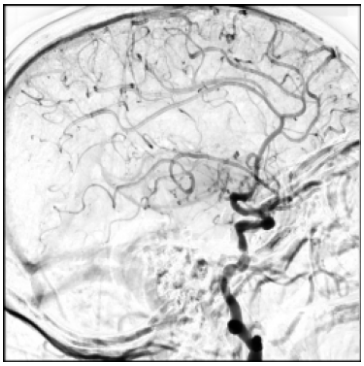


(b) DSA vein segmentation.

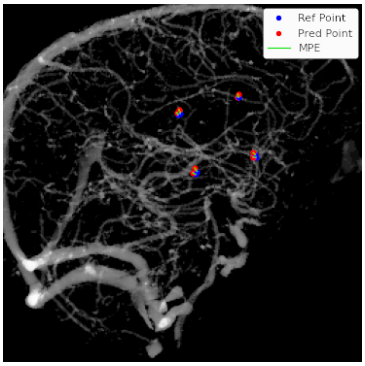


(c) Predicted initial pose.

(d) DSA full vein segmentation.

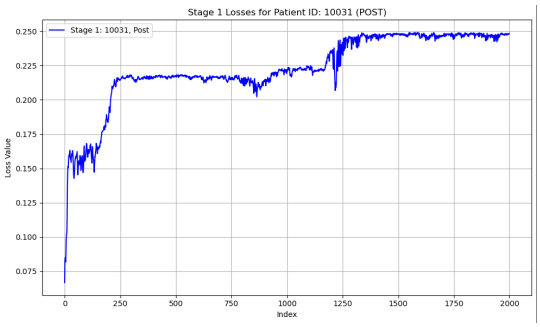


(e) DSA MinIP, used as input to second stage.

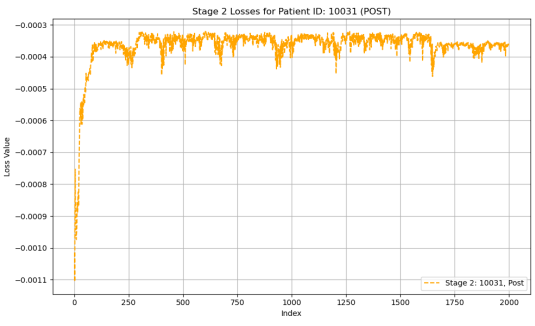



(f) Resulting pose from optimization.

Figure 3.7: Results from registering patient 10034. Distance to reference standard **before** registration: 17.1mm, distance **after**: 1.6mm. Distance to reference standard is plotted as the 8 points in green and blue.

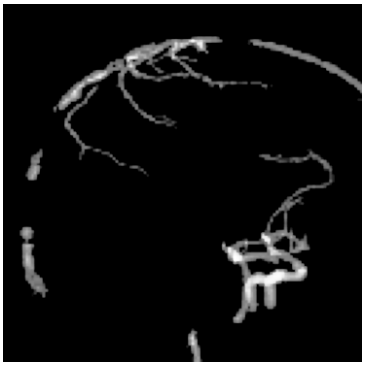



(a) Loss from first-stage optimization process.

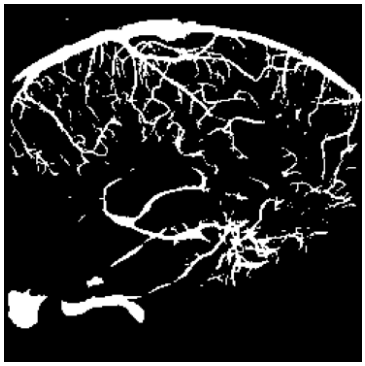



(b) Loss from second-stage optimization process.

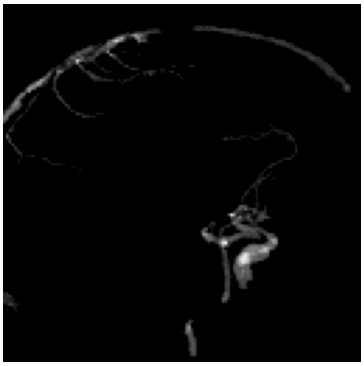
Figure 3.8: Losses from each optimization stage for patient 10034.



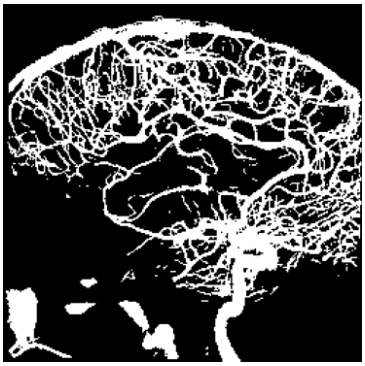
(a) Moving CTA.



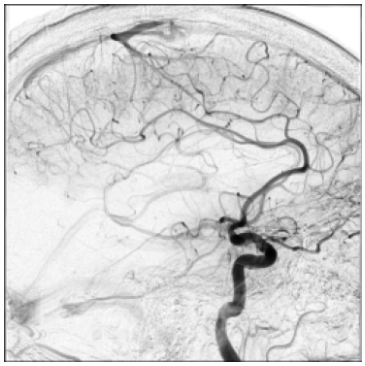
(b) DSA vein segmentation.



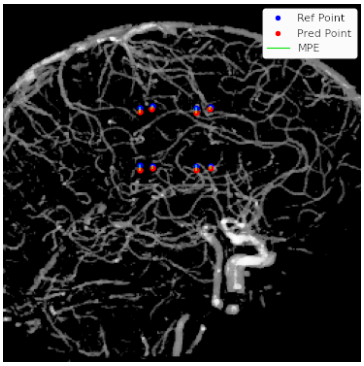
(c) Predicted initial pose.



(d) DSA full vein segmentation.

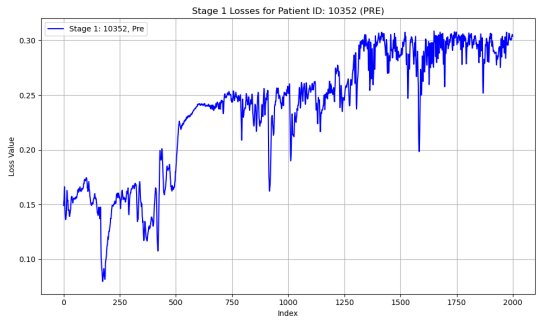


(e) DSA MinIP, used as input to second stage.

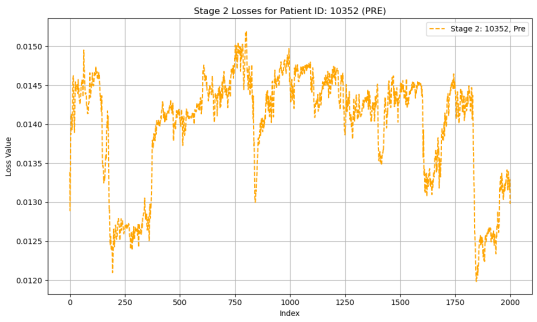



(f) Resulting pose from optimization.

Figure 3.9: Results from registering patient 10352. Distance to reference standard **before** registration: 17mm, distance **after**: 2.1mm. Distance to reference standard is plotted as the 8 points in green and blue.




(a) Loss from first-stage optimization process.




(b) Loss from second-stage optimization process.

Figure 3.10: Losses from each optimization stage for patient 10352 (PRE).

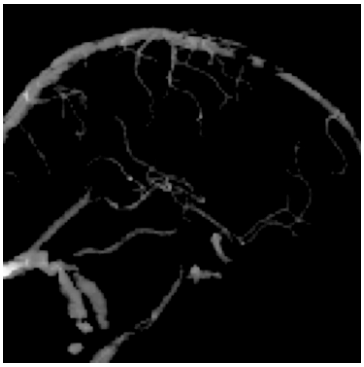
The resulting matrices are

$$\mathbf{T}_{\texttt{ref}} \qquad \mathbf{T}_{\texttt{pred}}$$

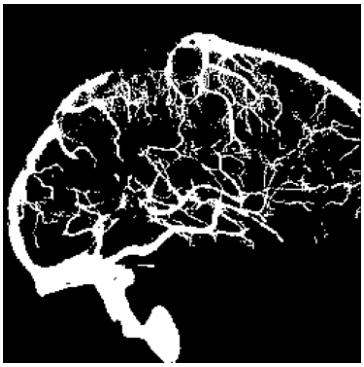
$$\begin{bmatrix} 0.2313 & 0.9727 & 0.0168 & -727.4096 \\ -0.9729 & 0.2314 & 0.0009 & -159.7554 \\ -0.0030 & -0.0166 & 0.9999 & 5.0820 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.2260 & 0.9739 & -0.0227 & -726.1982 \\ -0.9741 & 0.2261 & 0.0002 & -155.3700 \\ 0.0053 & 0.0221 & 0.9997 & -22.8050 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.5}$$

### 3.4.3. Registration Example: Patient R3166

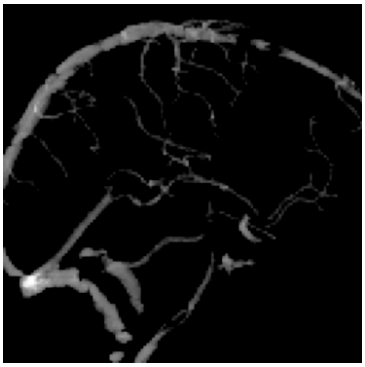
The DRRs and DSA images used are provided in Figure 3.11, and respective losses in Figure 3.12.



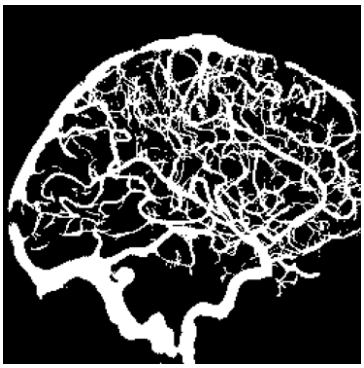
(a) Moving CTA.



(b) DSA vein segmentation.

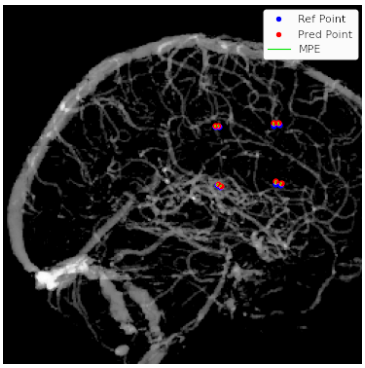


(c) Predicted initial pose.



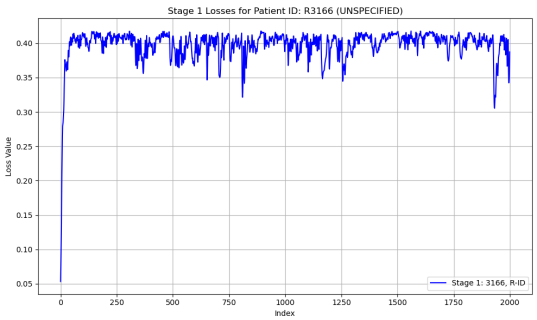
(d) DSA full vein segmentation.

(e) DSA MinIP, used as input to second stage.

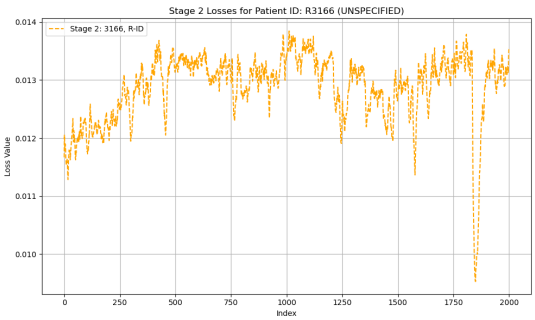



(f) Resulting pose from optimization.

Figure 3.11: Results from registering patient 3166. Distance to reference standard **before** registration: 19mm, distance **after**: 1mm. Distance to reference standard is plotted as the 8 points in green and blue.




(a) Loss from first-stage optimization process.




(b) Loss from second-stage optimization process.

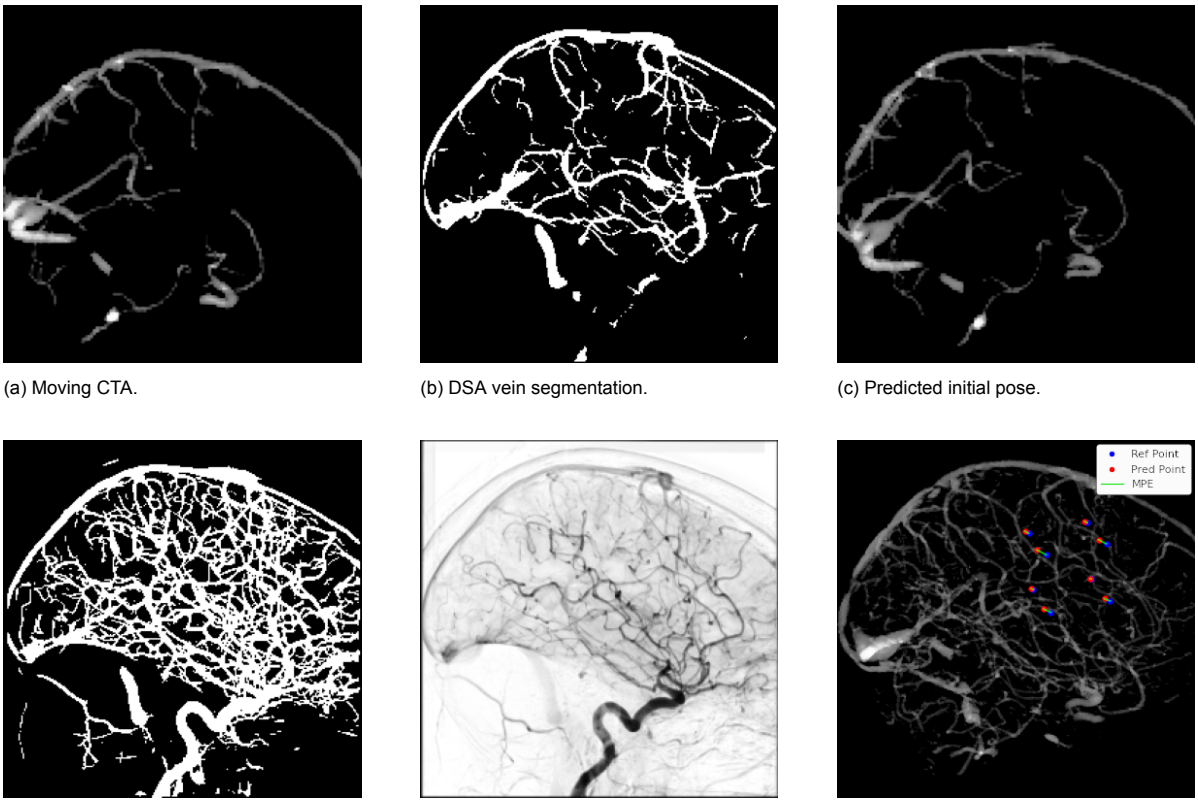
Figure 3.12: Losses from each optimization stage for patient 3166.

The resulting reference standard and predicted matrices are:

$$\underset{\mathbf{T}_{\text{ref}}}{\begin{bmatrix} 0.0344 & 0.9992 & 0.0226 & -777.8052 \\ -0.9991 & 0.0339 & 0.0238 & -41.4345 \\ 0.0230 & -0.0234 & 0.9995 & 19.3096 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \quad \underset{\mathbf{T}_{\text{pred}}}{\begin{bmatrix} 0.0381 & 0.9988 & 0.0291 & -779.1739 \\ -0.9982 & 0.0367 & 0.0472 & -44.1766 \\ 0.0461 & -0.0309 & 0.9985 & 25.1037 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \tag{3.6}$$

### 3.4.4. Registration Example: Patient 10031

The DRRs and DSA images used are provided in Figure 3.13, and respective losses in Figure 3.14.




(a) Moving CTA.

(b) DSA vein segmentation.

(c) Predicted initial pose.

(d) DSA full vein segmentation.

(e) DSA MinIP, used as input to second stage.

(f) Resulting pose from optimization.

Figure 3.13: Results from registering patient 10031. Distance to reference standard **before** registration: 22.1mm, distance **after**: 3.4mm. Distance to reference standard is plotted as the 8 points in green and blue.

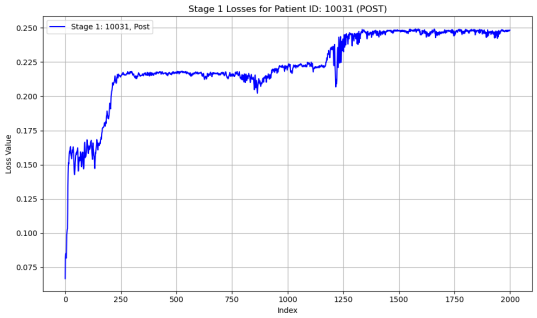
The resulting reference standard and predicted matrices are:

$$\underset{\mathbf{T}_{\text{ref}}}{\begin{bmatrix} 0.2714 & 0.8840 & 0.3805 & -650.0737 \\ -0.9591 & 0.2152 & 0.1840 & -172.4475 \\ 0.0807 & -0.4149 & 0.9063 & 304.8692 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \quad \underset{\mathbf{T}_{\text{pred}}}{\begin{bmatrix} 0.1852 & 0.9226 & 0.3385 & -676.3203 \\ -0.9767 & 0.1348 & 0.1668 & -109.8677 \\ 0.1082 & -0.3615 & 0.9261 & 262.5147 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \tag{3.7}$$

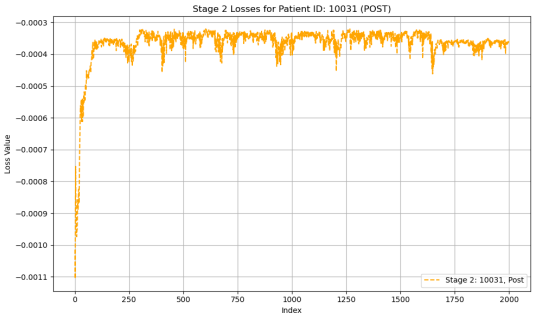
While the distance to the reference standard may seem large, 3.4mm in this case, the radiologist ranked the automatic registrations as consistently being better registered than the manual registrations, with both the three-stage and two-stage approach receiving a scale of 4.

### 3.4.5. Registration Example: Patient 10149 (POST)

The DRRs and DSA images used are provided in Figure 3.15, and respective losses in Figure 3.16.

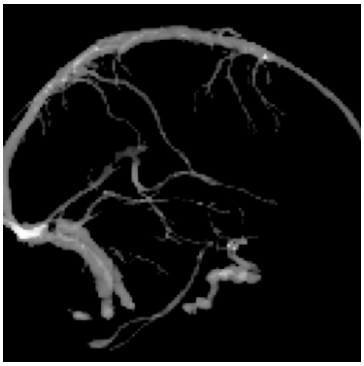



(a) Loss from first-stage optimization process.



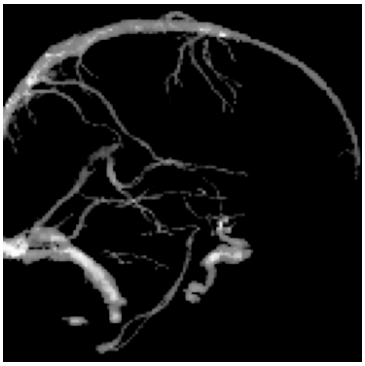

(b) Loss from second-stage optimization process.

Figure 3.14: Losses from each optimization stage for patient 10031.



(a) Moving CTA.

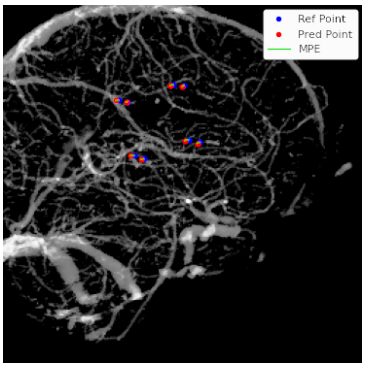
(b) DSA vein segmentation.



(c) Predicted initial pose.

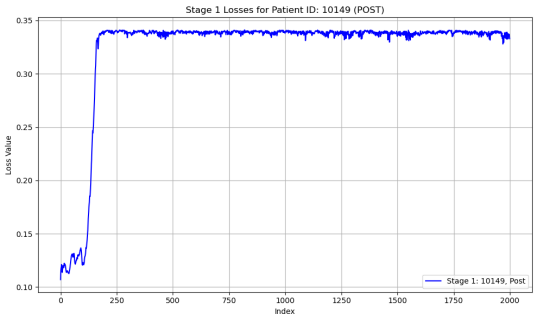
(d) DSA full vein segmentation.

(e) DSA MinIP, used as input to second stage.

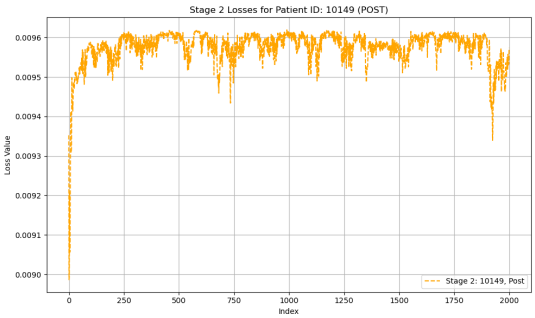



(f) Resulting pose from optimization.

Figure 3.15: Results from registering patient 10149 (post-EVT). Distance to reference standard **before** registration: 20.3mm, distance **after**: 1.7mm. Distance to reference standard is plotted as the 8 points in green and blue.




(a) Loss from first-stage optimization process.




(b) Loss from second-stage optimization process.

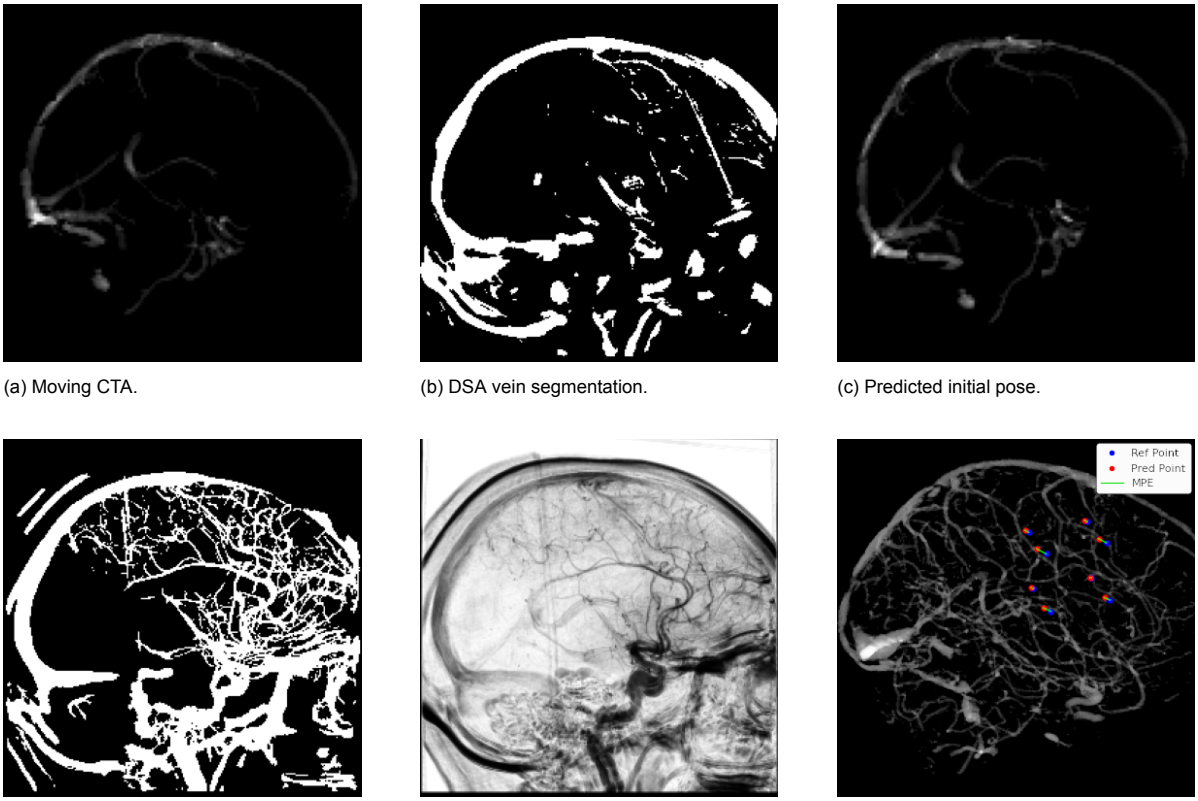
Figure 3.16: Losses from each optimization stage for patient 10031.

The resulting reference standard and predicted matrices are:

$$\underset{\mathbf{T}_{\texttt{ref}}}{\begin{bmatrix} 0.1539 & 0.9863 & 0.0599 & -749.4033 \\ -0.9565 & 0.1335 & 0.2594 & -90.1607 \\ 0.2478 & -0.0972 & 0.9639 & 59.1473 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \quad \underset{\mathbf{T}_{\texttt{pred}}}{\begin{bmatrix} 0.2093 & 0.9750 & 0.0746 & -730.6987 \\ -0.9446 & 0.1818 & 0.2733 & -121.5695 \\ 0.2529 & -0.1276 & 0.9590 & 80.3591 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \tag{3.8}$$

### 3.4.6. Registration Example: Patient 10320 (PRE)

The DRRs and DSA images used are provided in Figure 3.17, and respective losses in Figure 3.18.




(a) Moving CTA.

(b) DSA vein segmentation.

(c) Predicted initial pose.

(d) DSA full vein segmentation.

(e) DSA MinIP, used as input to second stage.

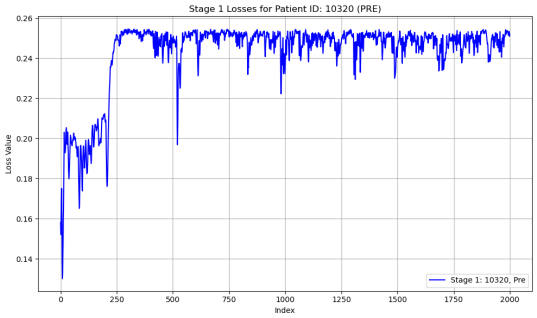
(f) Resulting pose from optimization.

Figure 3.17: Results from registering patient 10320. Distance to reference standard **before** registration: 10.7mm, distance **after**: 1.8mm. Distance to reference standard is plotted as the 8 points in green and blue.

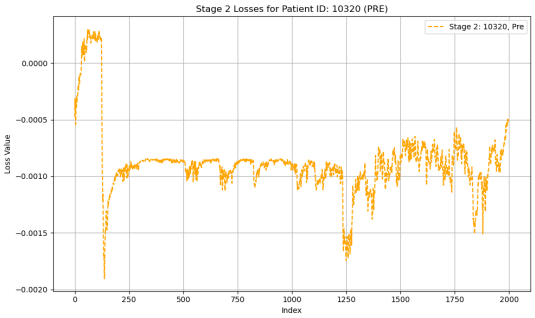
This patient makes for an interesting case due to the suboptimal quality of the DSA. The DSA image from this patient is performed pre-EVT. Due to the occlusion, this can result in many vessels not being visible in the DSA due to lack of blood flow, as illustrated in the DSA segmentation. However, despite the lack of rich vascular detail, the method is able to accurately register the CTA, suggesting that a few arteries can be sufficient for an accurate registration. We believe that this makes for important future research—identifying which vessels are crucial to registration may allow for a better qualification of whether image pairs are suitable for registration or not.

### 3.4.7. Failed Registration Example: Patient 10029

The registration below is an example of a failed case. This can be directly observed in Figure 3.19f, where the distance between the projected points is large. Interestingly enough, MPE distance before registration is not larger than in previous success cases. Associated losses are given in Figure 3.20.

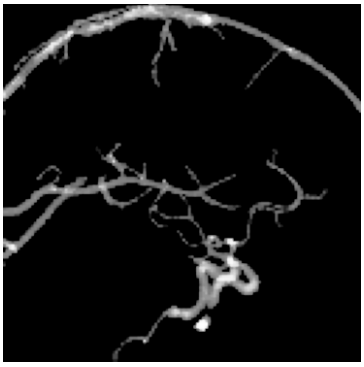



(a) Loss from first-stage optimization process.

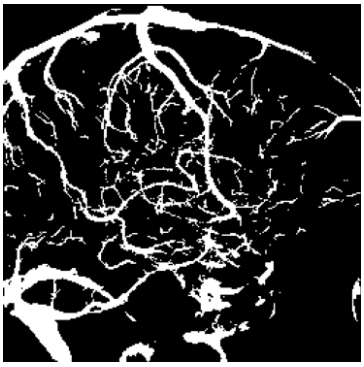



(b) Loss from second-stage optimization process.

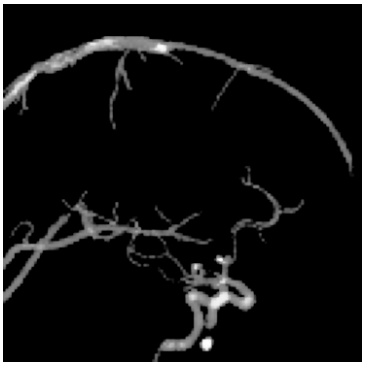
Figure 3.18: Losses from each optimization stage for patient 10029.



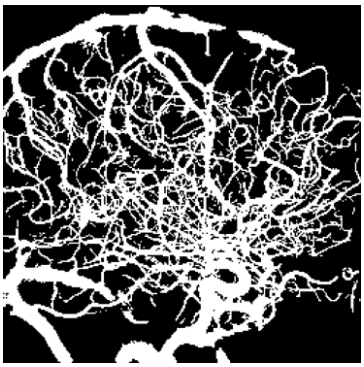
(a) Moving CTA.



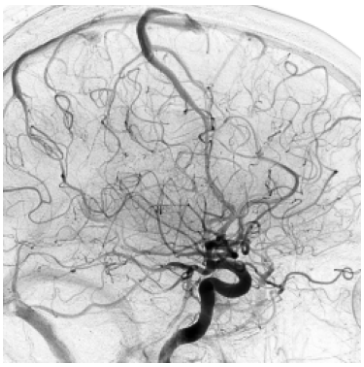
(b) DSA vein segmentation.



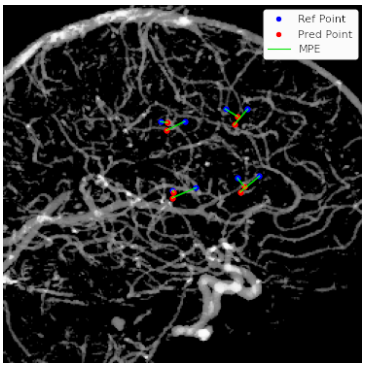
(c) Predicted initial pose.



(d) DSA full vein segmentation.

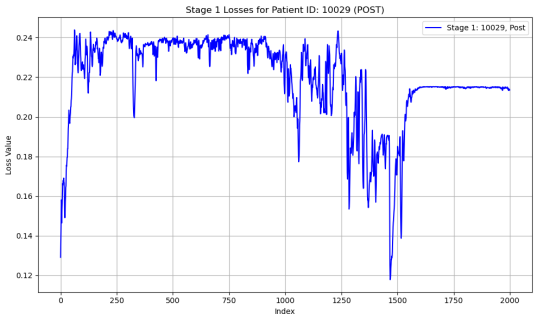


(e) DSA MinIP, used as input to second stage.

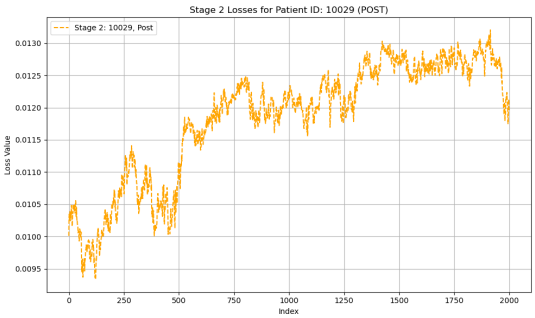



(f) Resulting pose from optimization.

Figure 3.19: Results from registering patient 10029. Distance to reference standard **before** registration: 16.6, distance **after**: 7.6mm. Distance to reference standard is plotted as the 8 points in green and blue.




(a) Loss from first-stage optimization process.

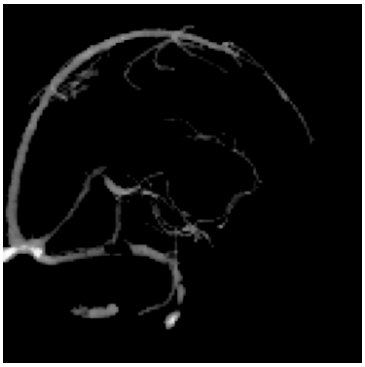



(b) Loss from second-stage optimization process.

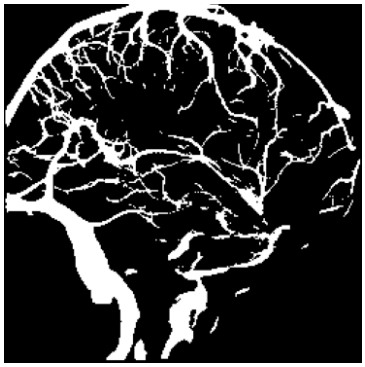
Figure 3.20: Losses from each optimization stage for patient 10029.

### 3.4.8. Failed Registration Example: Patient 10020

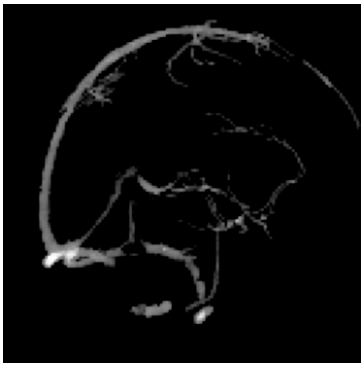
The registration below is an example of a failed case. This can be directly observed in Figure 3.21f, where the distance between the projected points is large. We consider this one of the *outlier* cases, due to the large rotation present in the CTA, which is most likely beyond the capture range of the proposed method. Associated losses are given in Figure 3.22.



(a) Moving CTA.

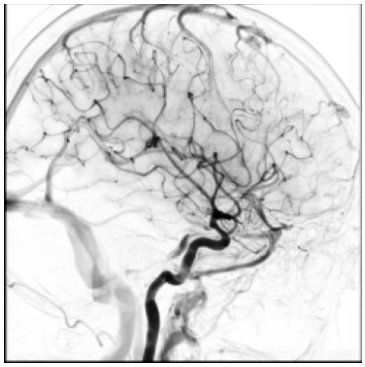


(b) DSA vein segmentation.

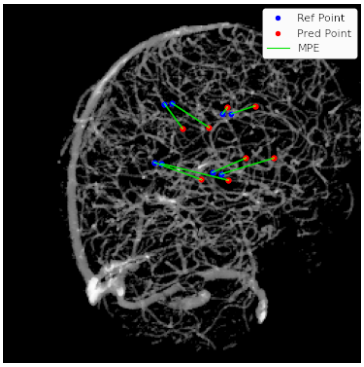


(c) Predicted initial pose.

(d) DSA full vein segmentation.

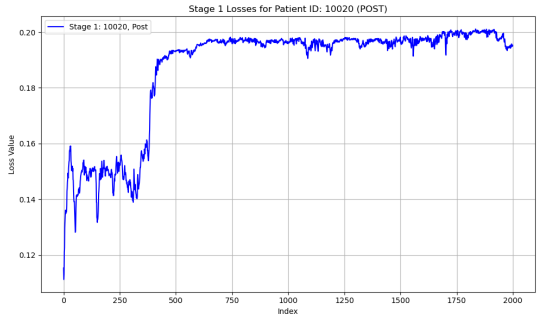


(e) DSA MinIP, used as input to second stage.

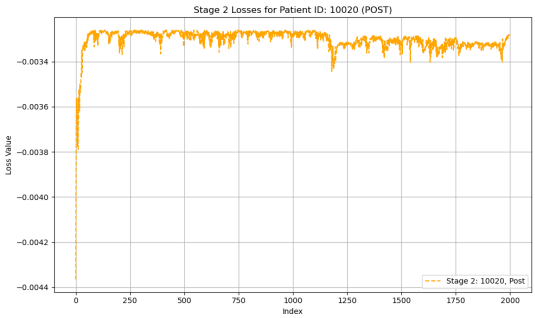



(f) Resulting pose from optimization.

Figure 3.21: Results from registering patient 10020. Distance to reference standard **before** registration: 47.9mm, distance **after**: 24.6mm. Distance to reference standard is plotted as the 8 points in green and blue.




(a) Loss from first-stage optimization process.




(b) Loss from second-stage optimization process.

Figure 3.22: Losses from each optimization stage for patient 10020.

## 3.5. Analysis of Initialization Poses

We hypothesized that an initial anchoring of the CTA based on the veins would be sufficient to 1) have a registration that is within the capture range of the optimization method, and 2) be sufficiently simple for a ResNet18 to learn. In this section, we plot DRRs generated from the initial pose predicteds from

the network, and overlay them with the ground truth, allowing for a visual inspection of whether or not the network is learning a pose that transforms the CTA to a position where it aligns with the venous structures. The simplest visual inspection is to overlay the predicted DRR over the top of the reference standard DRR, referred to as *after* in Figure 3.23. We also overlay the DRR rendered to its radiological pose with the reference standard pose, referred to as *before*. This side-by-side comparison allows for a visual inspection as to 1) how far the radiological pose was from the registered pose, 2) whether or not the predicted pose improved the alignment of the venous structures. A good initialization should result in the veins overlapping in the overlaid DRRs. In most success cases, the network seems to be performing an initial alignment based on the veins present in the CTA. This can be confirmed by identifying that in the DRR rendered according to the pose predicted by the network, there is a higher overlap in the venous structures.

It can also be observed that some initializations did not improve the initial alignment of the veins, such as patient 10020 (POST), in Figure 3.232. This is most likely due to the heavy rotation present in the CTA in its radiological pose. This patient is an example of a failed case in the full pipeline, one can observe that the patient's head is almost 45 degrees, making it somewhere between an anterior-posterior and lateral pose. Similarly, with patient 10029 (POST), the venous structures were in better alignment before initialization than after. In cases like this, we use a similarity measure as a proxy for alignment to gauge the improvement brought by the network. If the DRR rendered from the predicted pose results in a lower loss with the DSA than the DRR rendered from the radiological pose, the initial pose is discarded and the optimization process begins with the radiological pose as its initial pose, as defined in Equation 3.9.

$$\mathbf{T}_{\texttt{init}} = \begin{cases} \mathbf{T}_{\texttt{pred}}, & \text{if } \mathcal{L}(\text{DRR}_{\text{predicted}}, \text{DSA}) \geq \mathcal{L}(\text{DRR}_{\text{radiological}}, \text{DSA}) \\ \mathbf{T}_{\texttt{radiological}}, & \text{otherwise} \end{cases} \tag{3.9}$$

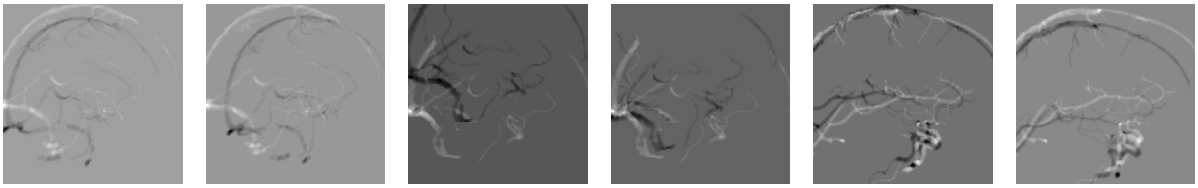
This check is required as the initialization that the network provides is only 'approximate'. It can be observed that, in some cases, the DRR rendered in its radiological pose is very close to the DRR rendered according to its reference standard pose, meaning there is very little movement to correct for. In cases like this, the initialization may result in a worse initial pose, thereby harming the optimization process.

## 3.6. CTA Vein Segmentation

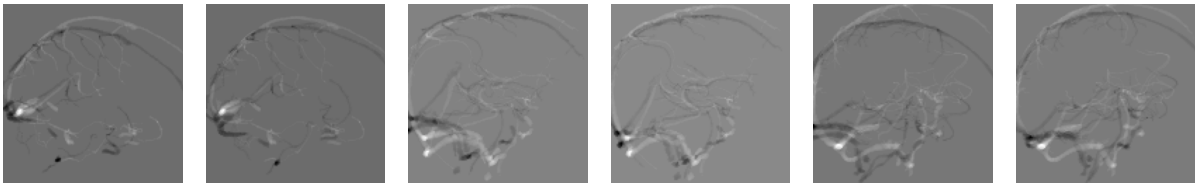
Examples of the CTA vein segmentations are given in Figure 3.23. Not all the vessels present in the DRRs are veins. While the morphological operations on the CTA were able to isolate the larger veins that surround the skull, other vessels, such as the sinuses, are still present. While this was sufficient to train the network, improvements to the training could be achieved by a more rigorous approach to segmenting the veins from the CTA, such as customizing the structuring element on a per-patient basis.

## 3.7. DSA Vein Segmentations

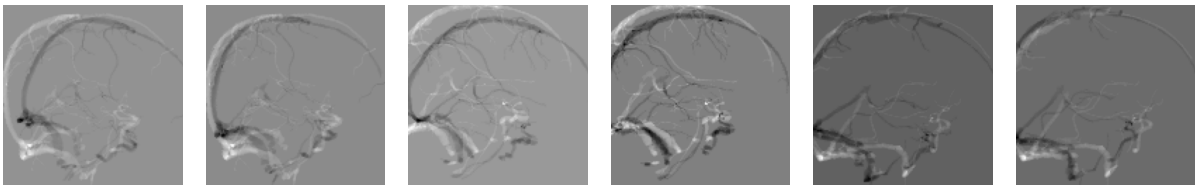
The 20 test DSA vein segmentations are provided in Figure 3.24.



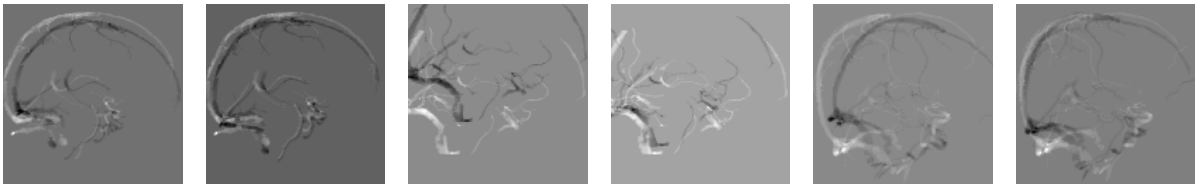
(1) 10020 (POST): Before (2) 10020 (POST): After (3) 10022 (POST): Before (4) 10022 (POST): After (5) 10029 (POST): Before (6) 10029 (POST): After



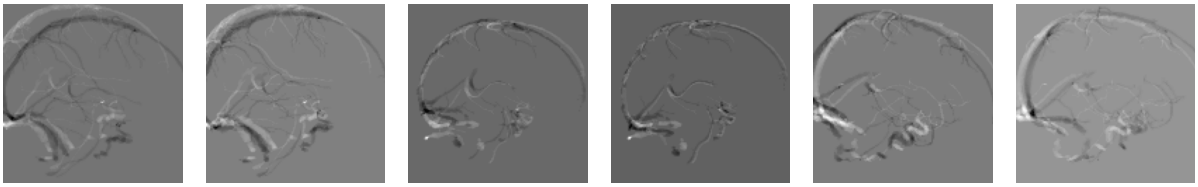
(7) 10031 (POST): Before (8) 10031 (POST): After (9) 10034 (POST): Before (10) 10034 (POST): After (11) 10050 (POST): Before (12) 10050 (POST): After



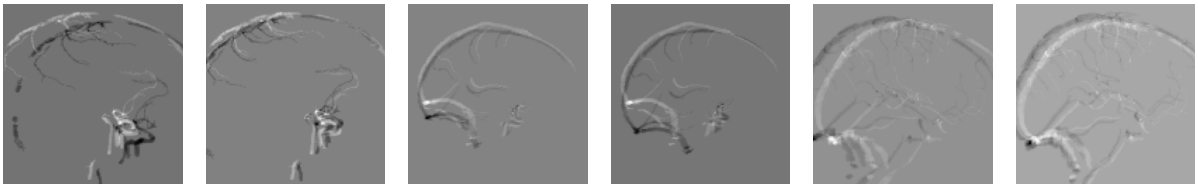
(13) 10051 (POST): Before (14) 10051 (POST): After (15) 10149 (POST): Before (16) 10149 (POST): After (17) 10153 (POST): Before (18) 10153 (POST): After



(19) 10320 (POST): Before (20) 10320 (POST): After (21) 10022 (PRE): Before (22) 10022 (PRE): After (23) 10051 (PRE): Before (24) 10051 (PRE): After



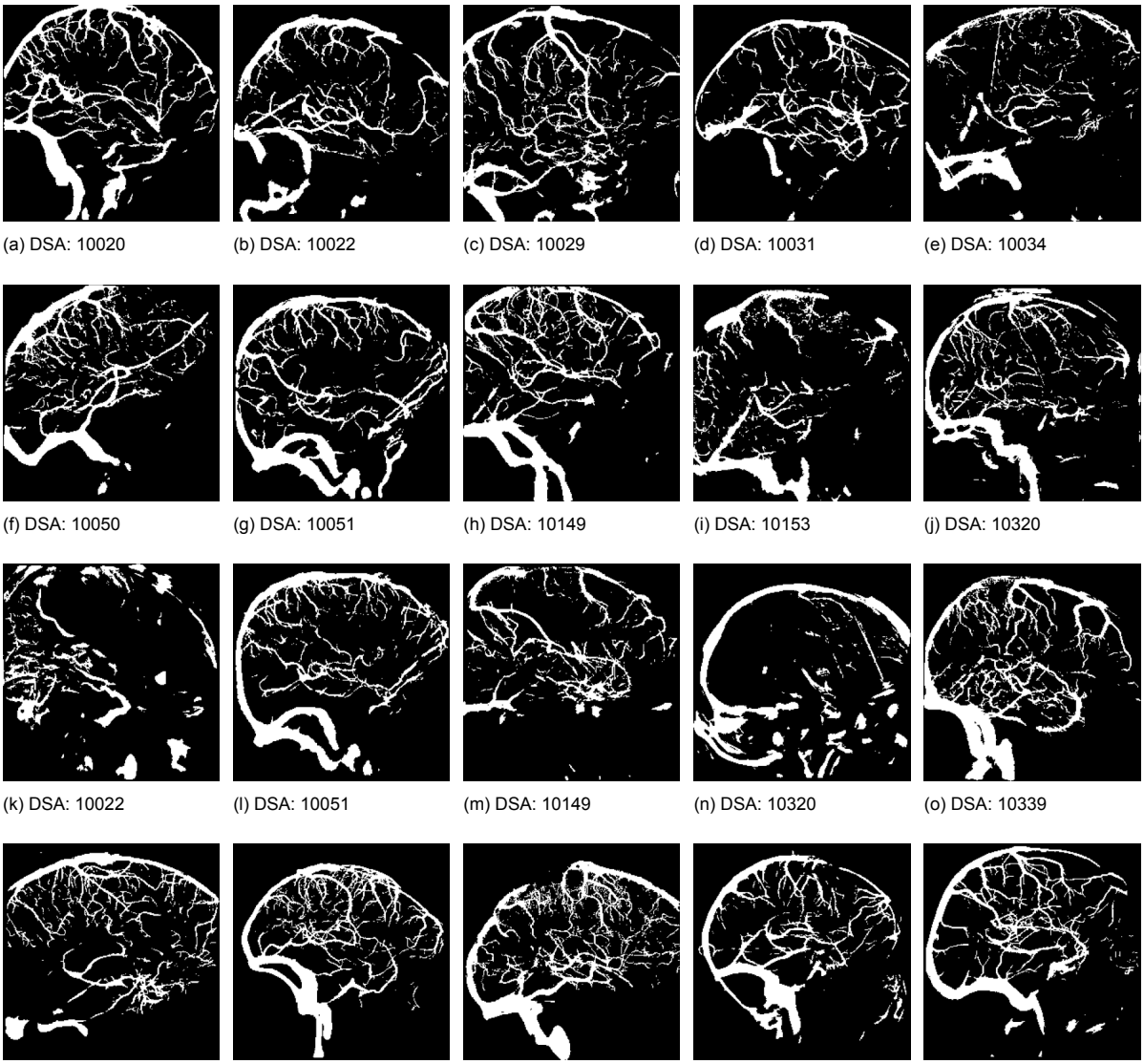
(25) 10149 (PRE): Before (26) 10149 (PRE): After (27) 10320 (PRE): Before (28) 10320 (PRE): After (29) 10339 (PRE): Before (30) 10339 (PRE): After



(31) 10352 (PRE): Before (32) 10352 (PRE): After (33) R2725 (NONE): Before (34) R2725 (NONE): After (35) R3166 (NONE): Before (36) R3166 (NONE): After

(37) R5249 (NONE): Before (38) R5249 (NONE): After (39) R5250 (NONE): Before (40) R5250 (NONE): After

Figure 3.23: Comparison of DRR rendered according to the radiological pose versus DRR rendered according to the predicted initial pose.



(a) DSA: 10020 (b) DSA: 10022 (c) DSA: 10029 (d) DSA: 10031 (e) DSA: 10034

(f) DSA: 10050 (g) DSA: 10051 (h) DSA: 10149 (i) DSA: 10153 (j) DSA: 10320

(k) DSA: 10022 (l) DSA: 10051 (m) DSA: 10149 (n) DSA: 10320 (o) DSA: 10339

(p) DSA: 10352 (q) DSA: R2725 (r) DSA: R3166 (s) DSA: R5249 (t) DSA: R5250

Figure 3.24: DSA vein segmentation for testset patients.

## 3.8. Matrix Construction and Coordinate System Conversion

In the research paper, the projection model is defined as:

$$\mathbf{p} = \mathbf{K} \cdot \mathbf{T}_{\text{man}} \cdot \mathbf{R}_{2w} \cdot \mathbf{T}_{\text{cent}} \cdot \mathbf{P}_w \quad , \tag{3.10}$$

where $\mathbf{R}_{2\text{w}}$ rotates the CTA into the 'face-up' position, and $\mathbf{T}_{\text{cent}}$ translates the CTA to the world coordinate center. These transformations are managed implicitly by TorchIO, the 3D medical imaging library used by DiffDRR, ensuring that a loaded CTA is aligned to the "canonical" pose, $\mathbf{T}_{\text{radiological}}$. $\mathbf{K}$ is constructed based on the C-arm parameters. However, the manual transformations, $\mathbf{T}_{\text{man}}$, used in the training data are provided in LPS coordinate systems, requiring conversion between coordinate systems in order to apply the manual registration matrix. This can be better understood via Figure 3.1. We outline the key transformations below.

### 3.8.1. Key Transformations

1 **Translation to World Center ($\mathbf{T}_{\text{cent}}$):**

This transformation centers the CTA in world coordinates. It is defined as:

$$\mathbf{T}_{\text{cent}} = \begin{bmatrix} 1 & 0 & 0 & -c_{\text{CTA}}[0] \\ 0 & 1 & 0 & -c_{\text{CTA}}[1] \\ 0 & 0 & 1 & -c_{\text{CTA}}[2] \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad ,$$

where $c_{\text{CTA}}$ is the center of the CTA in world coordinates.

2 **Rotation to Face-Up ($\mathbf{R}_{2\text{w}}$):** This ensures alignment of the CTA with the anatomical position. It is defined as:

$$\mathbf{R}_{2\text{w}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

3 **Manual Registration Transformation ($\mathbf{T}_{\text{man}}$):**

The manual transformation represents adjustments to align the CTA with the DSA. However, $\mathbf{T}_{\text{man}}$ provided in LPS must be converted to RAS+.

### 3.8.2. Conversion from LPS to RAS+ Coordinate System

To transform $\mathbf{T}_{\text{man}}$ into the RAS+ system, the following adjustments are applied:

1. **Negate Off-Diagonal Elements:** Flip necessary elements of the matrix to align the axes:

$$\begin{aligned} \mathbf{T}_{\text{man, adj}}[1,2] &= -\mathbf{T}_{\text{man}}[1,2], \\ \mathbf{T}_{\text{man, adj}}[2,1] &= -\mathbf{T}_{\text{man}}[2,1], \\ \mathbf{T}_{\text{man, adj}}[0,2] &= -\mathbf{T}_{\text{man}}[0,2], \\ \mathbf{T}_{\text{man, adj}}[2,0] &= -\mathbf{T}_{\text{man}}[2,0]. \end{aligned}$$

2. **Negate Translation Components:** Adjust translation components to match the RAS+ convention:

$$\mathbf{T}_{\text{man, adj}}[0,3] = -\mathbf{T}_{\text{man}}[0,3], \quad \mathbf{T}_{\text{man, adj}}[1,3] = -\mathbf{T}_{\text{man}}[1,3].$$

3. **Apply Face-Up Transformation:** Transform the adjusted matrix using $\mathbf{R}_{2\text{w}}$:

$$\mathbf{F}_{\text{tman}} = \mathbf{R}_{2\text{w}} \cdot \mathbf{T}_{\text{man, adj}}.$$

4. **Reverse the Alignment:** Use the inverse of the face-up transformation to return to the canonical pose:

$$\mathbf{T}_{\text{man, final}} = \mathbf{F}_{\text{tman}} \cdot \mathbf{R}_{2\text{w}}^{-1}.$$

### 3.8.3. Compose with Radiological Pose

Finally, the manually adjusted transformation matrix is composed with the radiological pose:

$$\mathbf{T}_{\texttt{final}} = \mathbf{T}_{\texttt{pose}} \cdot \mathbf{T}^{-1}_{\texttt{man, final}}.$$

### 3.8.4. Python Implementation

The Python implementation of the described process is as follows:

```
T_man_copy = T_man.clone()

# Apply necessary adjustments to R
T_man_copy[1, 2] = -T_man_copy[1, 2]
T_man_copy[2, 1] = -T_man_copy[2, 1]
T_man_copy[0, 2] = -T_man_copy[0, 2]
T_man_copy[2, 0] = -T_man_copy[2, 0]

# Negate t_x and t_y in t
T_man_copy[0, 3] = -T_man_copy[0, 3]
T_man_copy[1, 3] = -T_man_copy[1, 3]

# Convert to apply-transform coordinate system
F_tman = torch.matmul(R_faceup, T_man_copy).to(self.device)
F_inv = torch.inverse(R_faceup).to(self.device)
manual = torch.matmul(F_tman, F_inv).to(self.device)
manual = RigidTransform(manual)

# Apply the final pose
final_pose = pose.to(self.device).compose(manual.inverse().to(self.device))
```

Further details on the coordinate systems used for the registration matrices can be found in the repository containing the tool used in order to perform the manual registrations, available here: https://gitlab.com/radiology/igit/q-maestro/manual-2d-3d-registration/-/tree/main

## 3.9. Human Assessment Tool

As discussed in the research paper, a neurointerventional radiologist was queried in order to evaluate the results from the automatic registration tool. To this end, we designed an interactive Python tool that allowed the radiologist to process the images in order to assess the quality of each registration. A screenshot of the tool provided to the radiologist is given in Figure 3.25.

The tool has a variety of controls presumed to be relevant for evaluation by a radiologist. The radiologist was queried beforehand with a proposed set of controls in order to ensure the tool would be fit for the radiologist to assess each registration. The *Frame* control allows the radiologist to choose which frame in the DSA sequence is rendered. The *Use MinIP* button computes the MinIP of the DSA, where the start and end frame are given in the *Start Frame* and *End Frame* sliders. The *Alpha* control corresponds to the $\alpha$ coefficient in the linear interpolation used to overlay the images. A *Color Tint* button was added which adds a red tint to the CTA render. This was necessary due to the difficulty of seeing the arteries overlaid in both modalities. The red tint helps to visualize which vessels in the CTA overlap with the DSA. The *Brightness* control then allows to adjust the brightness of the CTA.

For the 14 success cases patients, we asked the radiologist to compare the results from the manual registrations versus the two-stage results, the manual registration versus the three-stage results, and the two-stage results versus the three-stage results. For each patient this resulted in 3 comparisons. The radiologist was presented with two registrations, as presented in Figure 3.25. The radiologist ranked whether the left or right registration was better on a scale of 1 to 5:

- **Scale 1**: left definitely better than right
- **Scale 2**: left is better than right

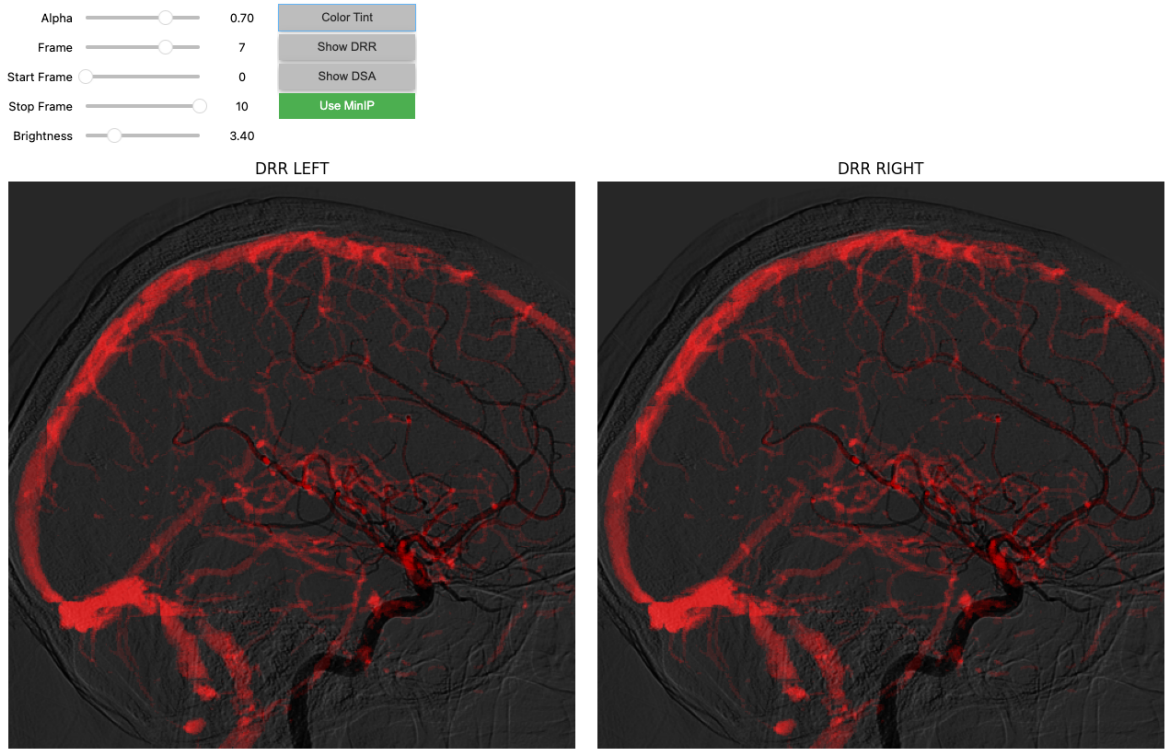



Figure 3.25: Illustration of the evaluation tool used to rank the registrations.

- **Scale 3**: left is the same as right
- **Scale 4**: right is better than left
- **Scale 5**: right definitely better than left

Each method was randomly given on the left or right, meaning there was no predictability in which method was on which side. The results from the radiologist are given in Table 3.1. A statistical interpretation of the results is given in the research paper.

Table 3.1: Summary of results from the radiologist evaluation.

| ID | Left | Right | Scale (1-5) | Best | Patient Info |
|---|---|---|---|---|---|
| 1 | TWO_STAGE | THREE_STAGE | 4 | THREE_STAGE | R3166 (UNSPECIFIED) |
| 2 | TWO_STAGE | MANUAL | 1 | TWO_STAGE | R3166 (UNSPECIFIED) |
| 3 | THREE_STAGE | MANUAL | 1 | THREE_STAGE | R3166 (UNSPECIFIED) |
| 4 | THREE_STAGE | TWO_STAGE | 3 | = | R5249 (UNSPECIFIED) |
| 5 | MANUAL | TWO_STAGE | 3 | = | R5249 (UNSPECIFIED) |
| 6 | MANUAL | THREE_STAGE | 4 | THREE_STAGE | R5249 (UNSPECIFIED) |
| 7 | MANUAL | TWO_STAGE | 4 | TWO_STAGE | 10031 (POST) |
| 8 | THREE_STAGE | MANUAL | 2 | THREE_STAGE | 10031 (POST) |
| 9 | TWO_STAGE | THREE_STAGE | 3 | = | 10031 (POST) |
| 10 | TWO_STAGE | MANUAL | 2 | TWO_STAGE | 10034 (POST) |
| 11 | THREE_STAGE | TWO_STAGE | 5 | TWO_STAGE | 10034 (POST) |
| 12 | MANUAL | THREE_STAGE | 2 | MANUAL | 10034 (POST) |
| 13 | THREE_STAGE | MANUAL | 2 | THREE_STAGE | 10339 (PRE) |
| 14 | TWO_STAGE | THREE_STAGE | 3 | = | 10339 (PRE) |
| 15 | MANUAL | TWO_STAGE | 4 | TWO_STAGE | 10339 (PRE) |
| 16 | MANUAL | THREE_STAGE | 4 | THREE_STAGE | 10149 (PRE) |
| 17 | TWO_STAGE | MANUAL | 2 | TWO_STAGE | 10149 (PRE) |
| 18 | THREE_STAGE | TWO_STAGE | 3 | = | 10149 (PRE) |
| 19 | TWO_STAGE | THREE_STAGE | 4 | THREE_STAGE | 10352 (PRE) |
| 20 | MANUAL | TWO_STAGE | 2 | MANUAL | 10352 (PRE) |
| 21 | THREE_STAGE | MANUAL | 3 | = | 10352 (PRE) |
| 22 | MANUAL | THREE_STAGE | 4 | THREE_STAGE | 10050 (POST) |
| 23 | THREE_STAGE | TWO_STAGE | 3 | = | 10050 (POST) |
| 24 | TWO_STAGE | MANUAL | 4 | MANUAL | 10050 (POST) |
| 25 | THREE_STAGE | TWO_STAGE | 1 | THREE_STAGE | 10153 (POST) |
| 26 | MANUAL | THREE_STAGE | 2 | MANUAL | 10153 (POST) |
| 27 | TWO_STAGE | MANUAL | 5 | MANUAL | 10153 (POST) |
| 28 | TWO_STAGE | MANUAL | 1 | TWO_STAGE | 10320 (PRE) |
| 29 | MANUAL | THREE_STAGE | 5 | THREE_STAGE | 10320 (PRE) |
| 30 | THREE_STAGE | TWO_STAGE | 3 | = | 10320 (PRE) |
| 31 | MANUAL | TWO_STAGE | 1 | MANUAL | 10320 (POST) |
| 32 | THREE_STAGE | MANUAL | 3 | = | 10320 (POST) |
| 33 | TWO_STAGE | THREE_STAGE | 3 | = | 10320 (POST) |
| 34 | THREE_STAGE | MANUAL | 1 | THREE_STAGE | 10149 (POST) |
| 35 | TWO_STAGE | THREE_STAGE | 3 | = | 10149 (POST) |
| 36 | MANUAL | TWO_STAGE | 5 | TWO_STAGE | 10149 (POST) |
| 37 | TWO_STAGE | THREE_STAGE | 3 | = | 10051 (POST) |
| 38 | THREE_STAGE | MANUAL | 1 | THREE_STAGE | 10051 (POST) |
| 39 | MANUAL | TWO_STAGE | 5 | TWO_STAGE | 10051 (POST) |
| 40 | MANUAL | THREE_STAGE | 4 | THREE_STAGE | 10051 (PRE) |
| 41 | TWO_STAGE | MANUAL | 3 | = | 10051 (PRE) |
| 42 | THREE_STAGE | TWO_STAGE | 3 | = | 10051 (PRE) |

## 3.10. IPCAI 2025 Long Abstract

The following pages contain our approved long abstract submission to IPCAI 2025 as “late breaking research”. A short communication version is being prepared for submission IJCARS.

# Bibliography


[1] 3D Slicer Community. *Coordinate Systems - 3D Slicer Wiki*. Accessed: 2025-01-09. n.d. URL: https://www.slicer.org/wiki/Coordinate_systems.

[2] AHA. “Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association”. In: *Circulation* 135 (2017), e146–e603.

[3] Olvert A Berkhemer et al. “A randomized trial of intraarterial treatment for acute ischemic stroke”. In: *New England Journal of Medicine* 372.1 (2015), pp. 11–20.

[4] Bruce CV Campbell et al. “Endovascular therapy for ischemic stroke with perfusion-imaging selection”. In: *New England Journal of Medicine* 372.11 (2015), pp. 1009–1018.

[5] Valery L Feigin et al. “Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century”. In: *The lancet neurology* 2.1 (2003), pp. 43–53.

[6] Yabo Fu et al. “Deep learning in medical image registration: a review”. In: *Physics in Medicine & Biology* 65.20 (2020), 20TR01.

[7] Vivek Gopalakrishnan, Neel Dey, and Polina Golland. *Intraoperative 2D/3D Image Registration via Differentiable X-ray Rendering*. 2023. arXiv: 2312.06358 [cs.CV].

[8] Vivek Gopalakrishnan and Polina Golland. “Fast Auto-Differentiable Digitally Reconstructed Radiographs for Solving Inverse Problems in Intraoperative Imaging”. In: *Clinical Image-based Procedures: 11th International Workshop, CLIP 2022, Held in Conjunction with MICCAI 2022, Singapore, Proceedings*. Lecture Notes in Computer Science. Springer, 2022. DOI: https://doi.org/10.1007/978-3-031-23179-7_1.

[9] Mayank Goyal et al. “Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials”. In: *The Lancet* 387.10029 (2016), pp. 1723–1731.

[10] Mayank Goyal et al. “Randomized assessment of rapid endovascular treatment of ischemic stroke”. In: *New England Journal of Medicine* 372.11 (2015), pp. 1019–1030.

[11] Grant Haskins, Uwe Kruger, and Pingkun Yan. “Deep learning in medical image registration: a survey”. In: *Machine Vision and Applications* 31 (2020), pp. 1–18.

[12] John H Hipwell et al. “Intensity-based 2-D-3-D registration of cerebral angiograms”. In: *IEEE transactions on medical imaging* 22.11 (2003), pp. 1417–1426.

[13] Tudor G Jovin et al. “Thrombectomy within 8 hours after symptom onset in ischemic stroke”. In: *New England Journal of Medicine* 372.24 (2015), pp. 2296–2306.

[14] David C. Lauzier and Akash P. Kansagra. “Thrombectomy in Acute Ischemic Stroke”. In: *New England Journal of Medicine* 386.14 (2022), pp. 1351–1351. DOI: 10.1056/NEJMicm2116727. eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMicm2116727. URL: https://www.nejm.org/doi/full/10.1056/NEJMicm2116727.

[15] Panagiotis Papanagiotou and George Ntaios. “Endovascular thrombectomy in acute ischemic stroke”. In: *Circulation: Cardiovascular Interventions* 11.1 (2018), e005362.

[16] Jeffrey L Saver et al. “Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke”. In: *New England Journal of Medicine* 372.24 (2015), pp. 2285–2295.

[17] H Bart Van der Worp and Jan van Gijn. “Acute ischemic stroke”. In: *New England Journal of Medicine* 357.6 (2007), pp. 572–579.