POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# NeuroGlue: Attentional Graph Based Mosaicking in Neurosurgery

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Author: **Anna Maria De Luca**

Student ID: 941435
Advisor: Prof. Elena De Momi
Co-advisors: Alessandro Casella
Academic Year: 2021-2022

# Abstract

The Surgical Microscope (SM) is the gold standard instrument in neurosurgery. The SM allows to visualise the surgical field and the anatomical details of the brain structures. On the other hand, the high magnifications provided by the SM cause a very limited field of view (FoV) that may lead to harmful operations on anatomical structures or a nearby organ, which will affect the surgical outcome.

In the opinion of Humanitas Hospital neurosurgeons, the illustrated limitations have an evident influence on Cerebral Cavernous Malformations (CCMs) removal and glioma resection.

A intra-operative system able to perform a real-time and broad exposure of the surgical theatre could be an effective tool to support neuro-surgeons in tumour or lesions localization and treatment.

The panorama can represent an important reference for surgeon as it allows him to observe the brain tissue details using the SM magnifications and at the same time, to consult a broader map of the operating field, without moving the SM.

The reconstruction and expansion process is obtained with a deep learning-based mosaicking framework, applied to some videos of a neurosurgical setting captured by a SM.

To the best of our knowledge, this work represents the first application of mosaicking on neurosurgical images.

The purpose of video mosaicking is to combine consecutive frames of a video sequence, in which each frame shows only an partial local view of the field of interest, in order to obtain a broader view of the same scene.

The classical mosaicking approach is characterized by the following four stages: keypoints detection and description; keypoints matching with outlier rejection; homography estimation; and image warping and blending.

The peculiar conditions of a neuro-surgical setting could affect the quality of the mosaicking result. Indeed the presence of regular patterns (for the blood vessel's structure),

viewpoint changes, illumination variations, and motion blur are relevant factors that make the classical mosaicking method not robust enough.

These issue introduces the need to find stable keypoints detectors and establish stronger matches between the keypoints of consecutive frames; essential characteristics for the correct homography matrix estimation and the consequent mosaicking accurate development.

The proposed architecture is called NeuroGlue and it is constituted by a Fully Convolutional Neural Network (named SuperPoint) for keypoints detection and description and by an Attentional Graph Neural Network for keypoints matching. The combination between the NeuroGlue layers stability and the domain adaptation performed during training allows to achieve a promising result, both in visual terms and also in terms of 5-frames Structural Similarity Metric (i.e. *SSIM*), standing out clearly from traditional algorithms (BRISK, ORB and SIFT) and also respect to the SuperPoint model pre-trained with the COCO dataset.

Mosaikcing for surgical theatre expansion answers to low visibility issue in neurosurgery with SM. Moreover its integration with navigation systems and preoperative images could minimize the inaccuracies for lesions or tumour localization and removal, improving the surgical outcome.

**Keywords:** Surgical microscope, Neurosurgery, Cerebral Cavernous Malformation, Glioma, Mosaicking, Deep learning.

# Sommario

Il microscopio chirurgico (SM) è lo strumento maggiormente utilizzato negli interventi di neurochirurgia.

Il microscopio chirurgico permette di visualizzare il campo operatorio e i dettagli anatomici delle strutture cerebrali esaminate. D'altra parte gli elevanti ingradimenti causano una riduzione sostanziale del campo visivo del chirurgo. Questa limitazione può risultare dannosa per le strutture anatomiche e gli organi coinvolti, influenzando l'esito dell'intervento chirurgico.

Secondo l'opinione di alcuni neurochirurghi dell'ospedale Humanitas di Milano, queste limitazioni hanno un particolare effetto sulla rimozione di Malformazioni Caveronse Cerebrali (CCM), note anche come Angiomi, e sulla resezione di gliomi.

Un sistema intra-operatorio capace di eseguire un'espansione del campo chirurgico in tempo reale potrebbe essere un efficiace strumento di supporto per il chirurgo nel trattamento e nella più agevole localizzazione di tumori e lesioni cerebrali.

Il panorama può rappresentare un importarte riferimento per il chirurgo; in quanto egli può sia osservare i dettagli del tessuto cerebrale grazie agli ingrandimenti del microscopio e contemporaneamente consultare una mappa più ampia dell'intero campo chirurgico, senza dover spostare il microscopio.

La ricostruzione del campo chirurgico si realizza tramite una rete di apprendimento per il mosaicking che riceve come input una sequenza di fotogrammi estratti da un video eseguito con microscopio.

Al meglio delle nostre conoscenze, questo lavoro rappresenta il primo tentativo di applicazione della tecnica del mosaicking a immagini neurochirurgiche.

Il processo di mosiacking mira a combinare una sequenza di immagini dove ognuna è caratterizzata da una visione parziale del campo di interesse. L'obbiettivo è quello di ottenere una rappresentazione più ampia della scena, la quale prende il nome di mosaico.

La metodologia classica del mosaicking comprende quattro fasi: rilevamento e descrizione

di punti chiave, identificazione delle corrispondenze tra i punti chiavi di fotogrammi consecutivi con eliminazione degli outlier, stima dell'omografia e infine trasformazione e unione delle immagini.

Le condizioni peculiari di un ambiente neurochirurgico rappresentano un ostacolo rilevante per la corretta esecuzione del mosaicking. Infatti la presenza di schemi ripetitivi (dati dalla struttura dei vasi sanguigni), variazioni di illuminazione e sfocatura da movimento nelle immagini, sono fattori che rendono i metodi tradizionali non sufficientemente robusti.

Questa problematica introduce il bisogno di utilizzare rilevatori e descrittori di punti chiave stabili e di calcolare corrispondenze più robuste tra i punti chiave di immagini consecutive. Queste sono caratteristiche essenziali per eseguire una corretta stima dell'omografia e il conseguente sviluppo accurato del mosaico.

L'architettura proposta prende il nome di NeuroGlue ed è costituita da una Rete Neurale Convoluzionale (chiamata SuperPoint) per il rilevamento e la descrizione dei punti chiave, e da una Rete Neurale a Grafo Attenzionale per la fase di calcolo delle corrispondenze. La combinazione tra la stabilità di NeuroGlue, data dalla struttura della rete, e la procedura di adattamento al dominio eseguita durante la fase di allenamento, permette di ottenere un risultato promettente, osservabile sia visivamente ma anche in termini di metrica di similarità strutturale calcolata ogni 5 fotogrammi (indicata come *SSIM*). Dunque il metodo proposto si distingue chiaramente rispetto agli algoritmi tradizionali (BRISK, ORB e SIFT) e anche in relazione alla rete neurale SuperPoint, pre-allenata con il dataset COCO.

La tecnica di mosaicking applicata all'espansione del campo operatorio risponde adeguatamente ai problemi relativi alla scarsa visibilità in neurochirurgia con microscopio. Inoltre la sua integrazione con i sistemi di navigazione e con le immagini preoperatorie potrebbe minimizzare le inaccuratezze riscontrate nella localizzazione e nel trattamento di lesioni cerebrali e tumori, migliorando l'esito dell'intervento chirurgico.

**Parole chiave:** Microscopio chirurgico, Neurochirurgia, Malformazione caveronsa cerebrala, Glioma, Mosaicking, Deep learning.

# Contents

# 1 | Introduction

## 1.1. Clinical context

Cerebral Cavernous Malformations (CCMs), also called Angiomas or Cerebral Cavernous Angiomas, are vascular lesions consisting of clusters of abnormally dilated blood vessels, also called caverns. These caverns have a typical raspberry appearance and they leak due to defects in the endothelial cells and in other structural components, required for vessel wall integrity. Lesion size is variable, ranging from microscopic to a few inches in diameter. [1]

The CCM appearance is caused by loss-of-function mutations to one of three CCM genes known as CCM1/KRIT1, CCM2/malcavernin or CCM3/PDCD10.

Angiomas occur in both sporadic and familial forms. Patients with familial CCMs typically have multiple malformations, with a correspondingly higher risk of complications. In contrast, the sporadic form is characterised by only a single lesion. In the familial CCM, even if all the endothelial cells present the gene mutation, the malformations are only found in a few localised regions of the brain microcirculation. Furthermore, it has been shown that, for human sporadic CCM, only a small fraction of endothelial cells has a mutation for the CCM genes. [2] These differences can be observed in Fig. 1.1. [3].

Subjects may be asymptomatic (quiescent state) or present a wide variety of symptoms including seizures, intracranial haemorrhages, or focal neurological deficits. [1]

The underlying pathology is reflected radiographically on Magnetic Resonance Imaging (MRI). [4]

The prevalence of CCM is not exactly known because the diagnosis can be made only with brain imaging or autopsy. Estimates from autopsy studies and MRI analysis suggest a frequency of 0.16% to 0.9%. [5]

The standard treatment is the neurosurgical excision, so the surgical removal of these lesions. [1] A complete resection of the lesion is mandatory because the presence of residuals increases the risk of haemorrhage and seizures.[6]

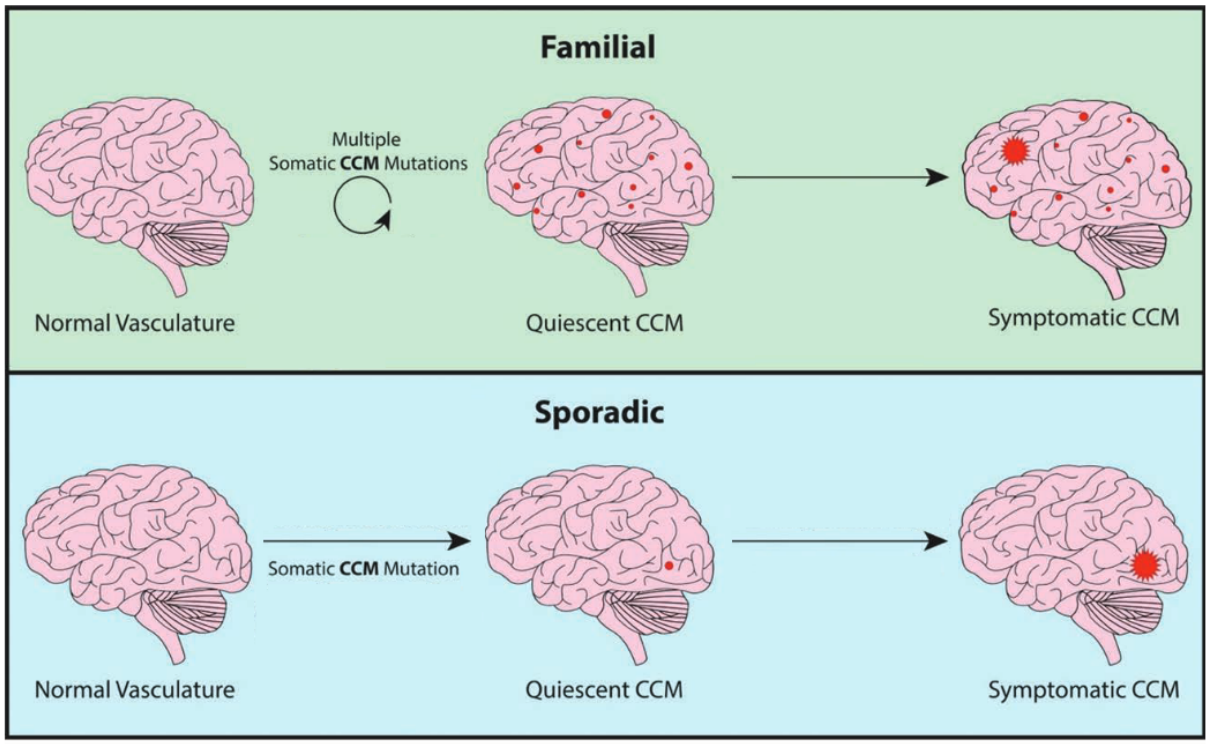


Figure 1.1: This image illustrates the difference between the familial and the sporadic CCM. The upper part of the image describes the familial CCM, characterized by several lesions. The down part shows the sporadic CCM, that presents only one cavern. The symptoms occur when the lesion size increases.

Glioma is one of the adults' most common primary malignant brain tumors. It can occur anywhere in the central nervous system but primarily in the brain.

They typically arise from glial tissue or precursor cells and develop into astro-cytoma, oligodendroglioma, ependymoma, or oligoastrocytoma.

According to the World Health Organization (WHO) classification, gliomas are categorized into four grades, among which grade 1 and grade 2 gliomas indicate a lower risk ones, and grade 3 and grade 4 gliomas indicate higher risk glioma. [7]

Glioblastoma is the most frequent grade 4 glioma, whose average age-adjusted incidence rate is 3.2 per $100,000$ population, as reported in [8]. In Fig. 1.2 an example of glioblastoma is illustrated with the corresponding post-operative situation which shows some small residuals. [9].

The presentation of a patient with glioma can vary greatly depending on the size and location of the tumour and on the anatomic structures of the involved nervous tissue. Also, for this reason, the glioma individuation is problematic; the common way, indicated

by the WHO, to confirm a diagnosis is the histologic analysis. [10]

The standard of care for gliomas includes maximal resection followed by concomitant radiotherapy and chemotherapy, especially in the 30 days after surgery. [7]

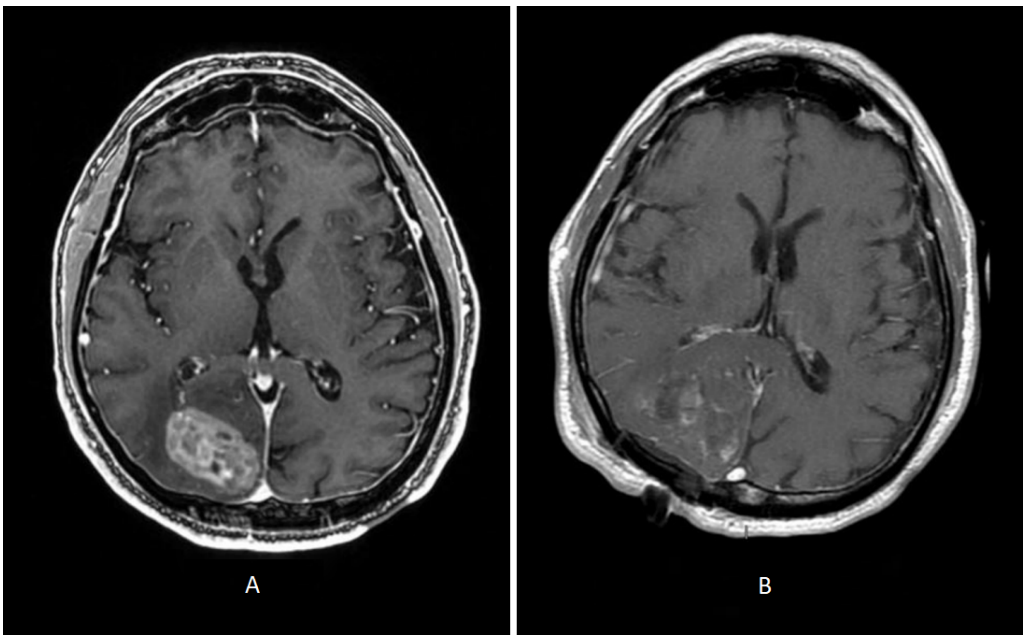


Figure 1.2: The figure shows the pre- (A) and immediate post-operative (at 24 hours) axial MRI (B). In the pre-operative image, an example of glioblastoma is present. In the post-operative image there is a minimal residual of the tumor.

The surgical resection is associated to several risks depending on the position and on the severity of the tumour. Most of the glioma surgery entails a standard craniotomy approach for its complete removal. The correct tumour localization and its overall view during the surgical treatment are necessary to reach a successful outcome. [11]

For the investigated surgical procedures and in general in the neurosurgery scenario, the surgical microscope (SM) is involved. [12]

The SM is the gold standard instrument in neurosurgery. [13] Indeed it allows to visualise the surgical field and the anatomical details of the brain structures, with the correct illumination. On the other hand, the high magnifications provided by the SM cause a very limited field of view (FoV) that may lead to harmful operations on anatomical structures or a nearby organ, which will affect the surgical outcome, reduce organ preservation, or even cause life-threatening consequences. [13, 14]

During the surgical operations intra-operative navigation systems are needed to detect the tumor or lesion position. They provide the rigid body transformation from the coordinate

system of the preoperative image to the coordinate system of the patient during surgery. [15] However registration errors due to this navigation systems may occur. In particular, the brain shift, which is a nonrigid brain deformation caused by craniotomy, reduces the alignment accuracy. [16]

In the opinion of Humanitas Hospital neurosurgeons, the illustrated limitations have an evident influence on Cerebral Cavernous Malformations (CCMs) removal and glioma resection.

An example of the real surgical environment is represented in Fig. 1.3 and a schematic scenario is shown in Fig. 1.4.

Therefore a reduced FoV could prevent the surgeon from having a complete view of the brain tumour or the lesion, and the navigation system errors may affect their localization. These are relevant factors that make neurosurgical procedures still challenging.

A intra-operative system able to perform a real-time and broad exposure of the surgical theatre could be an effective tool to support surgeons in tumors or lesions localization and treatment. Computer Assisted Intervention (CAI) is a powerful ally to deal with these types of challenge, especially through new developing techniques like Artificial Intelligence (AI) and Deep Learning (DL).

In particular mosaicking can be applied to extend the neurosurgical FoV, limited by the SM magnifications, by creating a panorama of the surgical environment.

This expansion of the operative field can be practised at any stage of the surgical intervention even if it was primarily tested to be applied in the early stage, generating a reconstruction of the brain superficial layers.

It is obtained performing a video registration of the FoV using the microscope, without the need to introduce a new sensor or a new device respect to the ones already present in the operating room. The panorama is visible in real-time and it can represent an important reference for the surgeon during the procedure.

Indeed the surgeon can work on the brain anatomy, observing each details, thanks to the high magnifications of the SM, and at the same time he is able to have a broader view of the entire scene, without further moving the microscope, storing the hand-held instruments or changing the magnifications during the procedure, reducing significantly the timing of surgery.

This tool could be involved in the surgical procedures thanks to its easy application for surgeons and because its employment is almost independent from surgeons actions or from

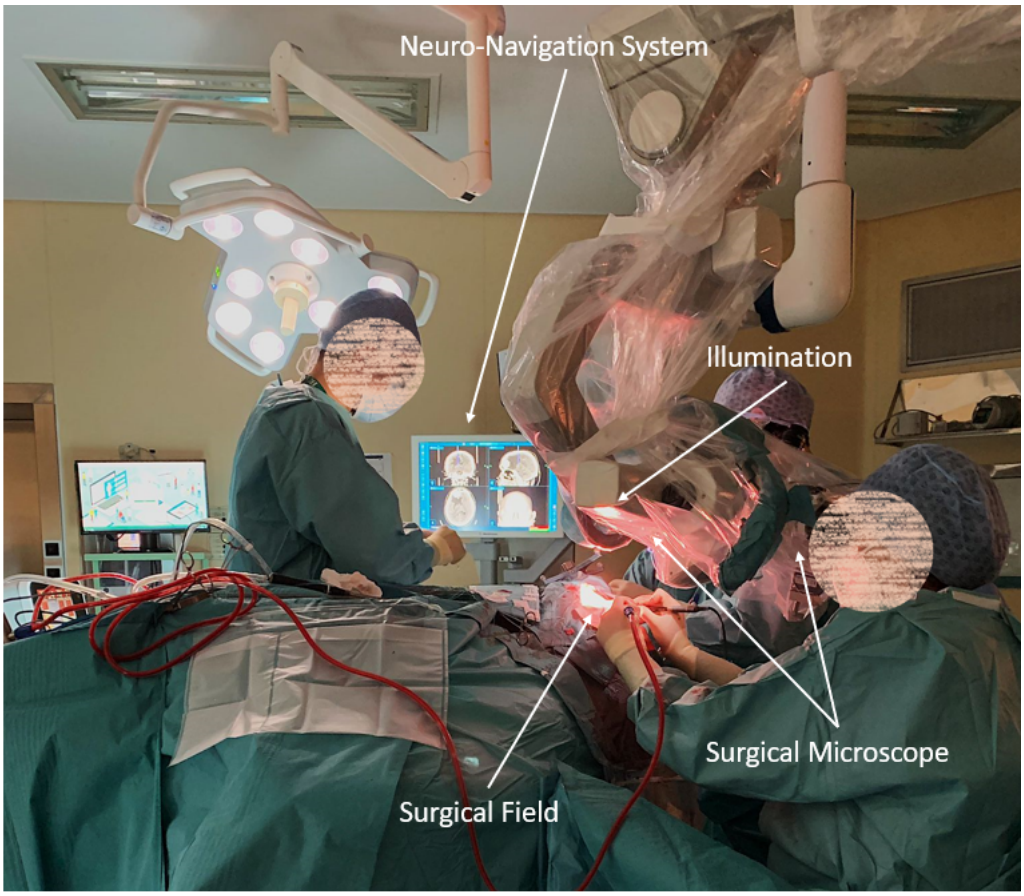


Figure 1.3: The figure illustrates a classical neurosurgical setting, characterized by the a) Surgical Microscope (SM), which provides an b) adequate illumination and is used by the surgeon to achieve a detailed view of the nervous tissues. In the image it is underlined the c) Surgical Field obtained with the SM and the d) Neuro-Navigation Systems, necessary for the lesion or tumor localization. The image was taken in a surgical room of Humanitas Research Hospital of Milano.

the occurrence of unexpected events such as a sudden movement of the SM.

These fast movements, caused for example by impacting the microscope during the registration of the surgical field, could affect the mosaicking outcome. [17] However the proposed method is able to eliminate the results obtained during the impact and restore the panorama previously recorded, without the need to start a new registration.

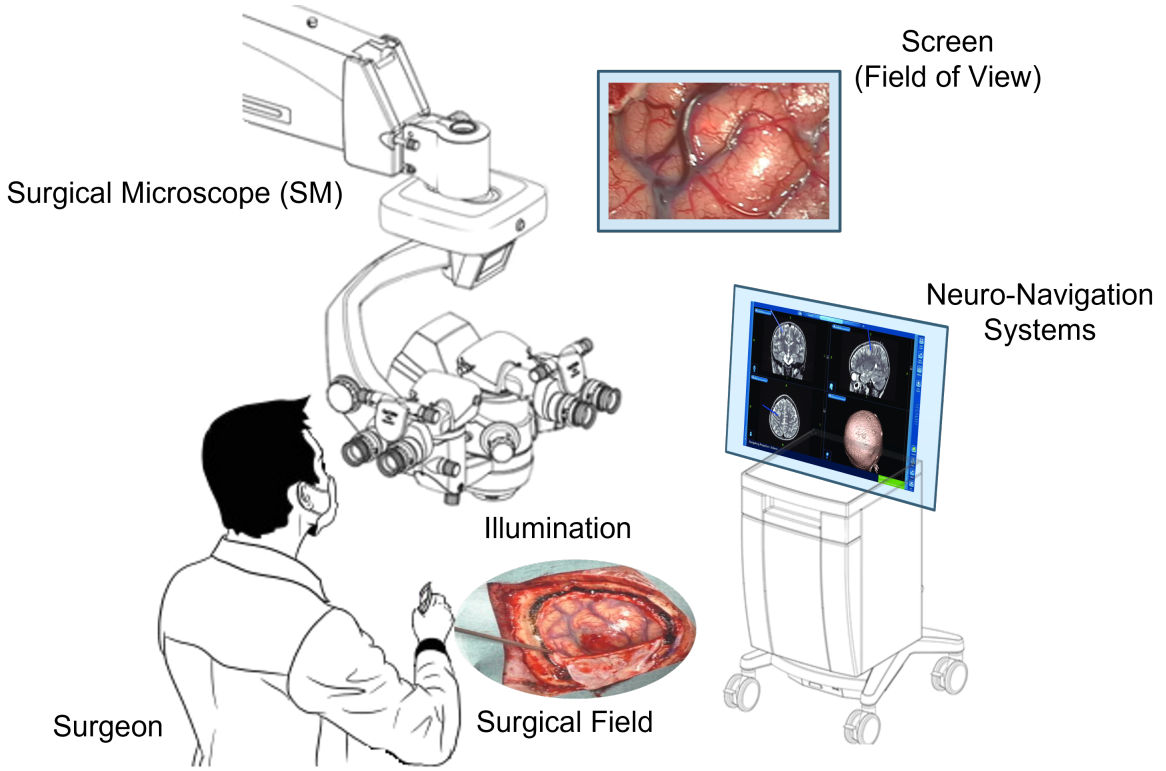


Figure 1.4: Schematic representation of the neurosurgical setting that puts into evidence the limited FoV, projected in the screen, the SM structure with the provided illumination and the navigation systems; taking inspiration from the real environment illustrated in Fig. 1.3.

## 1.2. Disclosure

The aim of this thesis is to implement a deep learning-based mosaicking framework obtained from videos of different neurosurgical settings, captured from a surgical microscope. The mosaicking approach is applied to expand the surgical field of view and support the surgeon during the operation, improving the surgical outcome.

This work was developed at *NearLab* in *Politecnico di Milano* in collaboration with *Humanitas Research Hospital of Milano* which made available the required technologies for this project.

# 2 | Literature Review

The purpose of video mosaicking is to combine different images, in which each one shows only a partial local view of the field of interest. It allows to obtain a broader view of the same scene. [17]

The classical mosaicking development is divided in four stages: i) Keypoint detection and description; ii) Keypoint matching and Outlier rejection; iii) Homography estimation; and iv) Image warping and blending, as it is illustrated in Fig. 2.1. In this process every step is essential for the correct execution of the next ones. [18]

The first step to be applied is the feature-detector, an algorithm that searches for keypoints into an image. Keypoints represent the most significant pixels of a considered image and can range from a single pixel to edges, contours, blobs, junctions and lines.

The detected keypoints are subsequently described in logically different ways based on unique patterns possessed by their neighboring pixels. This process is called feature description as it describes each keypoint by assigning it a distinctive identity which is represented with a vector that allows the effective keypoint recognition for matching. [18]

There are some keypoint detection algorithms that include also the description phase; instead other detectors have to be integrated and coupled with a descriptor algorithm.

Considering a frame pair, keypoints detection and description are applied in each image of the pair. Two sets of points (with their spatial coordinates) and descriptors vectors are obtained as output. These sets are characterized by a point-to-point correspondences, essential for matching. [19]

The next stage of the mosaicking process is the keypoint matching, which aims to establish correct feature correspondences from the two extracted points sets. [20]

Afterwards RANdom Sample Consensus (RANSAC) and Levenberg-Marquardt (LM) algorithm are jointly applied to estimate the homography. The objective of RANSAC is to select the optimal set of keypoints and filter out outliers that do not fit with a defined type of transformation. [21] Image warping and blending are then performed by com-

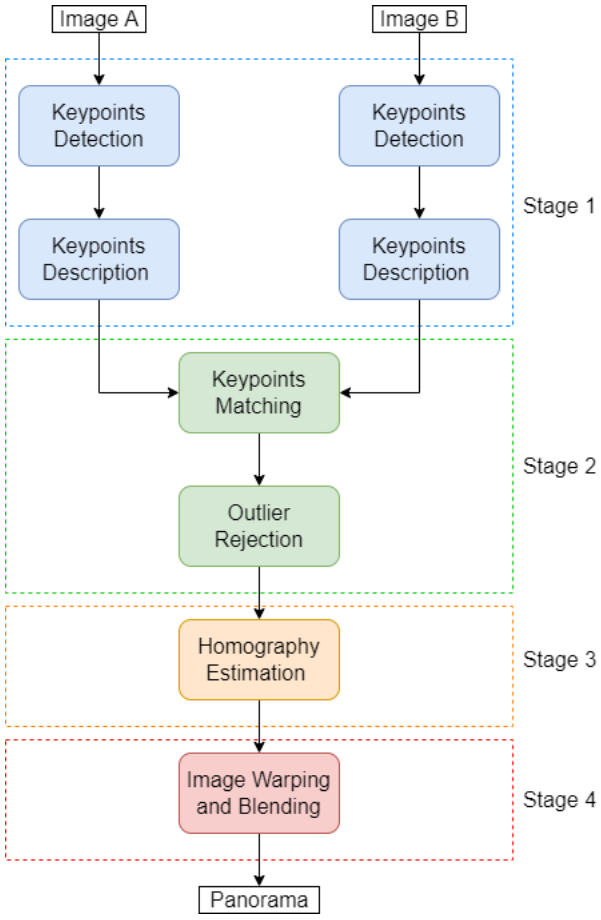


Figure 2.1: The classical four stages of image mosaicking are represented in this diagram:i) Keypoint detection and description; ii) Keypoint matching and Outlier rejection; iii) Homography estimation; and iv) Image warping and blending. In this figure, the steps are applied to an image pair (image A and B) in order to obtain a panorama image.

bining the two images, basing on the homography matrix previously computed until all matched feature points are aligned. In this way, a partial panorama is obtained. [22]

In the next sections some methods for keypoints detection description and matching, presented in literature, are described.

## 2.1. Keypoints Detection and Description

Since the very beginning of computer vision, feature detectors have occupied an important place in the research due to their numerous applications in such areas as object recognition, categorization, classification, robot localization and tracking, image matching and 3D reconstruction, image retrieval, registration, etc [22].

The research on feature detectors and descriptors is a fast growing area in image processing. The following section has been arranged chronologically to explain the gradual improvements in feature detection, as well as the evolution and limitations of comparative studies. [22]

The first corner detection algorithm was designed by Moravec in 1977 [23]. Harris and Stephens [24] revealed the limitations of this detector with their popular corner and edge detector (better known as Harris corner detector). It is based on the local auto-correlation function of the image for measuring the intensity differences between a patch and windows shifted in several directions.

Lucas and Kanade proposed a popular method, further developed by Tomasi and Kanade [25], often called the Kanade–Lucas–Tomasi feature tracker. This method allows to track the small motion of features in an image stream and it takes inspiration from the Harris and Stephens detector.

In the same decade other algorithms were designed, such as the one proposed by Heitger in [26] and the framework for low level feature extraction presented by Forstner in [27].

Shi and Tomasi [28] proposed a new detection metric (Good Features To Track or GFTT) based on the Harris operator, arguing that their model was a better choice. A completely new approach, called SUSAN, was also proposed in [29]. However, the illustrated operators of SUSAN fall short when rotation and scaling are involved. In 1998 a comprehensive review of a number of popular detectors is presented in [30]. The authors compare Harris corner detector to other algorithms previously mentioned in [26] and [27]. Later, Schmid et al. [31] revised the comparison method originally proposed, and conducted a number of qualitative tests. The test results indicated that the Harris operator (which is the basis of Harris corner detector) is the best among the compared methods.

Hall et al. [32] provided a definition of saliency, underlying that the most salient or discriminant image features are those that allow to distinguish one feature from others. An image feature that is present in only a single image or on a single object would allow to distinguish this image or object from all others. Moreover Hall et al. evaluated Harris, the method proposed in [33] regarding feature detection with automatic scale selection and Harris–Laplacian corner detector [34]. Harris–Laplacian is a combination of the Harris and Laplacian function for characteristic scale selection.

Driven by the need for a scale-invariant approach, Lowe [35] proposed one of the most popular feature detectors: the Scale Invariant Feature Transform or SIFT. SIFT is a combination of feature point detector and descriptor.

In this method, keypoints are extracted from the image in two steps. First, the image is repeatedly smoothed using Gaussian filters and subsampled to find images in smaller scales. In this way, an image pyramid is constructed with the reference image at the ground level (first octave). Second, keypoints are discovered in the $3 \times 3 \times 3$ neighborhood of any pixel at an intermediate level. These points are obtained from the image points where the Difference-of-Gaussians (DoG) values attain an extreme (minimum or maximum). This architecture gives scale and rotation invariance to the algorithm. [22, 36] Fig. 2.2 represents the SIFT detection structure.

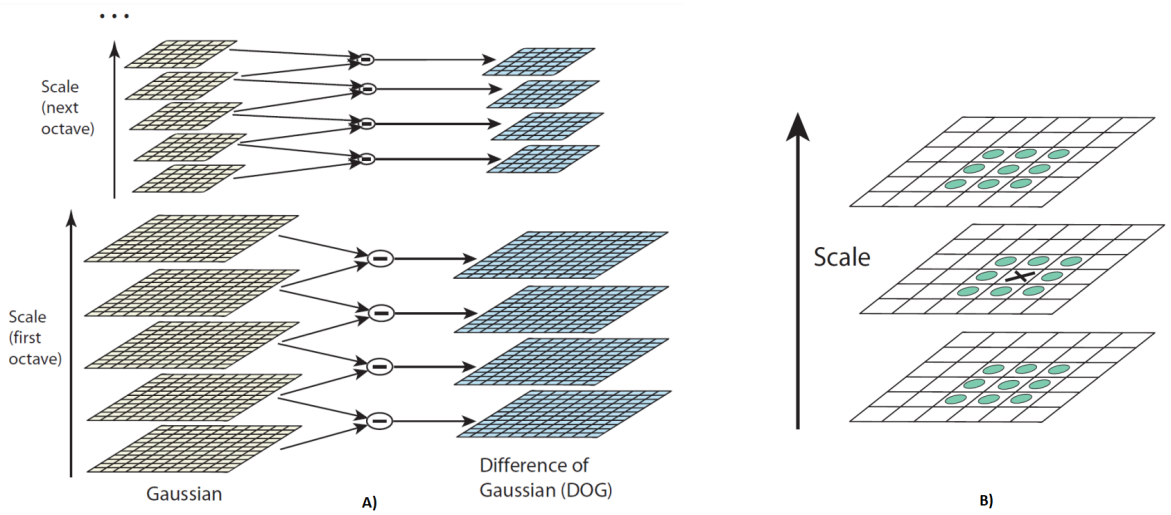


Figure 2.2: The figure shows the detection procedure of SIFT. A) is an example of Gaussian pyramid and of Difference-of-Gaussians (DoG). B) represents the search for extremes [37].

SIFT involves also the descriptor computation for the keypoints. The descriptor is related to the orientation of each detected key-point that is obtained using the point's gradients. In particular the gradients are rotated by the computed orientation and an histogram is created for each sub-region. [37] A common implementation is based on a $4 \times 4$ descriptor, as Fig. 2.3 illustrates, characterized by 8 different orientation vectors. [36, 37]

To resume, the SIFT descriptor is a position-dependent histogram of the local image gradient directions around the keypoints. [22] This feature detection algorithm is one of the most stable and precise between the ones illustrated in literature. [37]

Mikolajczyk and Schmid [38] compared the SIFT descriptor and a number of related descriptors, proving SIFT to be one of the best feature detectors based on the strength of its descriptor.

Harris and SIFT have been well explored and improved by several researchers [39–41].

Moreels and Perona [42] compared Harris, Hessian, and difference of Gaussian filters on images of 3D objects with different viewpoints, lighting variations, and scale variations. They differentiated between detectors and descriptors, using SIFT, PCA-SIFT, steerable filters, and shape context descriptors. They found a good match of detector and descriptor. In particular PCA-SIFT (Principal Component Analysis-SIFT) was developed to improve SIFT, projecting high-dimensional samples into a low-dimensional feature space and reducing the computational costs carried with Lowe's implementations. [43]

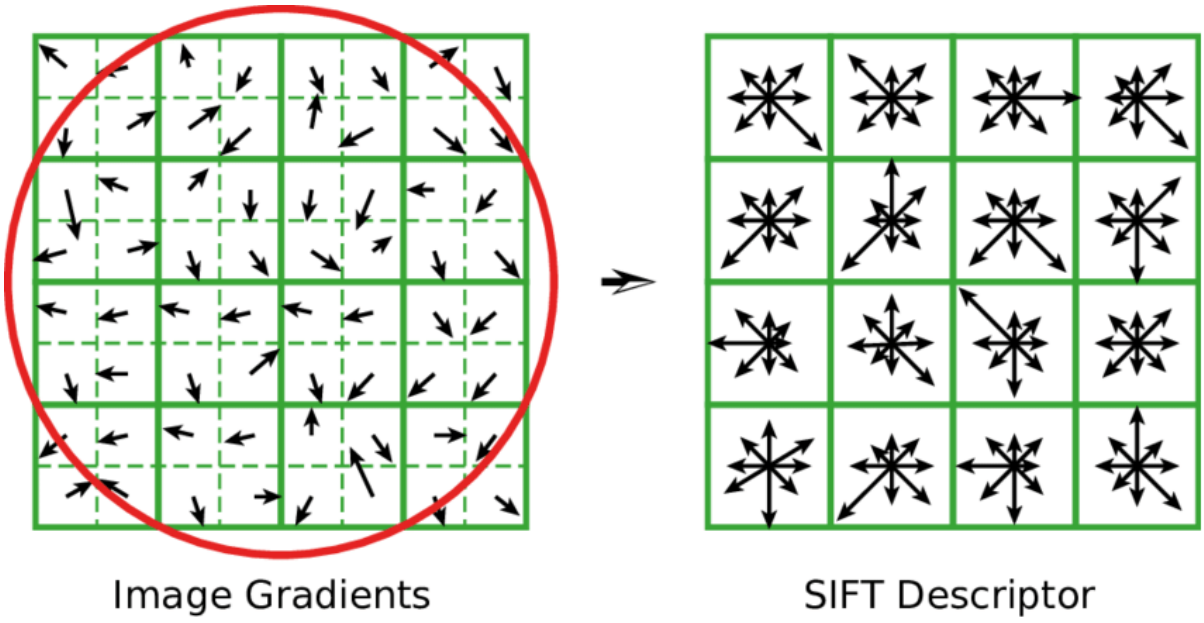


Figure 2.3: The figure represents the SIFT descriptor procedure [37].

In the following years several feature detectors were suggested taken inspiration from the SIFT algorithm.

Maximally Stable Extremal Regions (MSER), proposed by Matas et al. in [44], is a method able to detect blobs in images. The key points detection is based on a connectivity analysis, indeed it is performed by computing connected maximal and minimal intensity regions. A fast implementation of MSER can be found in [45].

Speeded Up Robust Features (SURF) is a scale and rotation-invariant feature detection and descriptor algorithm designed by Bay et al. [46]. The SURF detector is very similar to SIFT and it is found on each keypoint by orientation assignment and descriptor component analysis. [22]

Rosten and Drummond [47] described Features from Accelerated Segment Test (FAST) detector. FAST algorithm aims to find points of interest in an image and is specialized

in corner detection. The corner selection is done through some criteria which together create a decision tree in order to correctly classify all the corners. [22]

In particular FAST corner detector uses a circle of 16 pixels (a Bresenham circle of radius 3) to classify whether a candidate point $p$ is actually a corner. Each pixel in the circle is labeled from integer number 1 to 16 clockwise. If a set of $N$ contiguous pixels in the circle are all brighter than the intensity of candidate pixel $p$ (denoted by $I_p$) plus a threshold value $t$ or all darker than the intensity of candidate pixel $p$ minus threshold value $t$ , then $p$ is classified as corner. (Fig. 2.4) [48] Defining $S$, a set of $N$ contiguous pixels and $I_x$, the intensity of the pixels belonging to $S$, the conditions can be written as follows:

$$\begin{cases} Condition\ 1 : \forall x \in S, I_x > I_p + t \\ Condition\ 2 : \forall x \in S, I_x < I_p - t \end{cases} \tag{2.1}$$

There is a tradeoff of choosing $N$, the number of contiguous pixels and the threshold value $t$. On one hand the number of detected corner points should not be too many, on the other hand, the high performance should not be achieved by sacrificing computational efficiency. Usually $N$ is chosen as 12. [48]

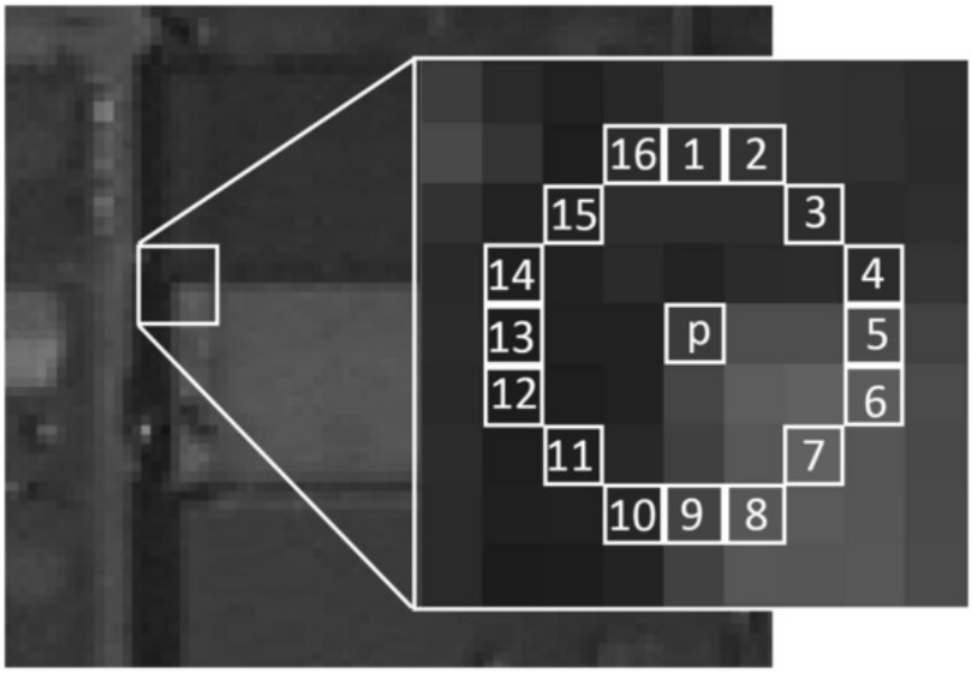


Figure 2.4: The figure represents the ORB detection mechanism, namely the FAST algorithm.[48].

Center Surround Extremas (CENSURE) is a detector proposed by Agrawal et al. [49]

based on two criteria: stability (persistence of features across viewpoint changes) and accuracy (consistency of feature localization across viewpoint changes). The feature detection is characterized by a method called Hessian–Laplacian.

Binary Robust Independent Elementary Features (BRIEF) was developed by Calonder et al. [50] in 2010. It is an attractive descriptor of binary strings that allows to extract descriptors from feature points for image matching. BRIEF is used a common descriptor for those detectors that don't have their own descriptors, such as FAST, CENSURE, and MSER.

It employs simple tests using intensity difference to create binary feature vectors that effectively describe key points in a pair of images. Before executing binary tests, images are smoothed using a Gaussian kernel at a pixel level, reducing noise sensitivity. The obtained binary strings with BRIEF only require between 128 and 512 bits, a relatively few number of bits compared with other state-of-the-art feature descriptors. The Hamming distance is employed for evaluating the ranking of descriptors, instead of the Euclidean distance, since it is easier to calculate. Although construction and matching for this descriptor is faster than other state-of-the-art ones, does not provide rotation and scale invariance, but tolerates just small amounts of rotations. [36]

Fusing together the FAST and the BRIEF algorithms the Oriented FAST and Rotated BRIEF (ORB) detector and descriptor is obtained. It was proposed by Rublee et al. [51] as a valid alternative to SURF and SIFT. The ORB descriptor modifies the FAST extractor adding an orientation component through first-order moments in a local patch. Then the BRIEF descriptor is computed on a rotated patch order to deal with the lack of rotational invariance. [52]

Another important feature detector and descriptor is the Binary Robust Invariant Scalable Key-points (BRISK), designed by Leutenegger et al. [53]. In this method, points of interest are first identified using a saliency criterion. Next, a sampling pattern is applied to the neighborhood of each of these detected key points to retrieve the orientation and so the descriptor.

It is based on a scalespace FAST method; in particular, it takes the input image $c_0$ to create a multiscale space with $n$ octaves $c_i$ and intra-octaves $d_i$, where $i$ goes from 0.1 to $n-1$ and normally $n=4$. To find the keypoints, a circular mask consisting of 8 points is used. If an adequate number of pixels is larger or smaller than the central one, it is taken as a candidate key-point. This process is performed in each octave and intra-octave separately to identify potential corner points [54].

The candidate points are compared with their neighbors in the same layer and in the upper $c_{i+1}$ and lower ones $c_{i-1}$, as shown in Fig. 2.5. Thus, points are evaluated on different planes and scales accordingly. After obtaining the keypoints at different scales, they are used for the construction of the descriptor.

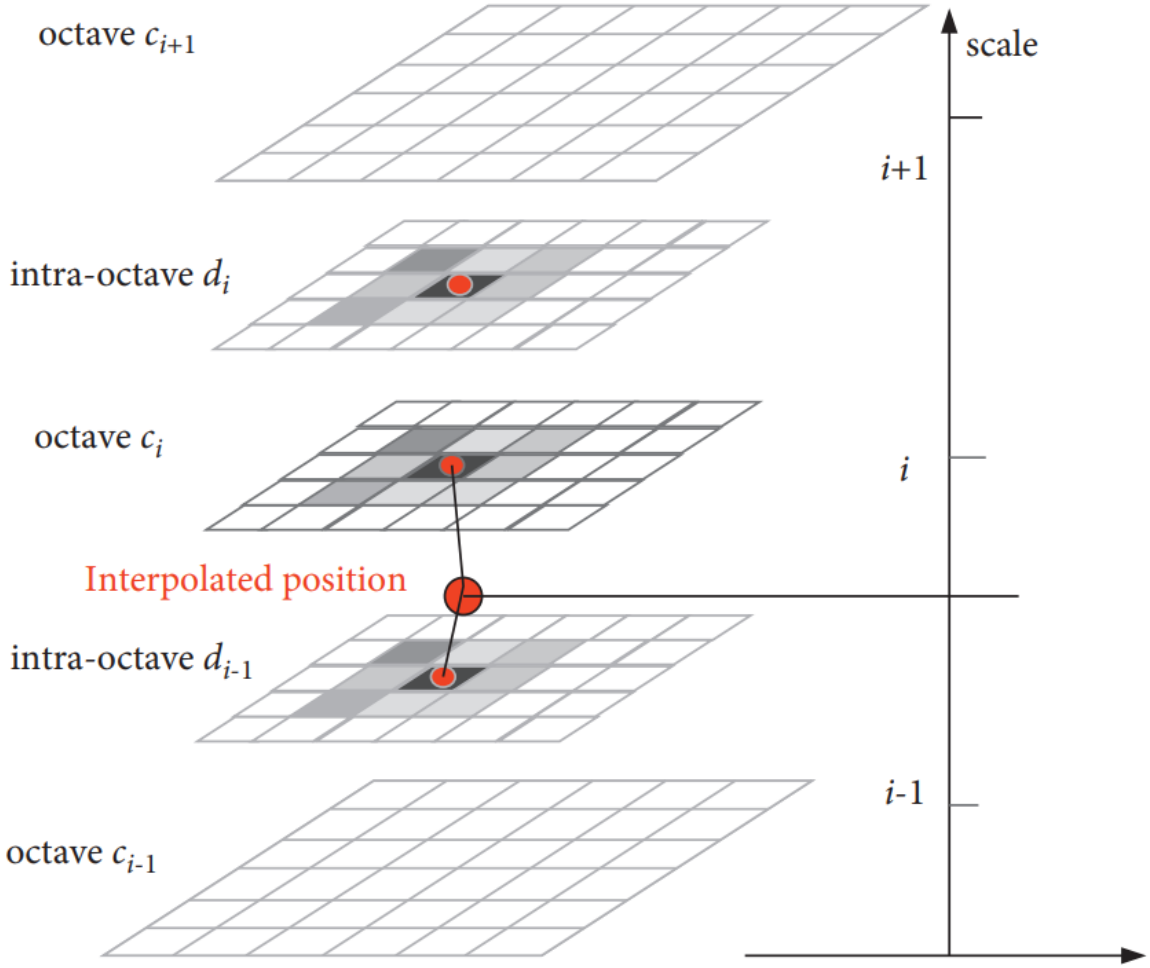


Figure 2.5: The figure represents the BRISK detection mechanism: a potential keypoint is identified in the octave by comparing 8 pixels of a neighbourhood $c_i$ as well as the corresponding patches of the immediately adjacent layers above $c_{i+1}$ and below $c_{i-1}$. [54].

Thus, on these $k$ points in the input image a sampling pattern of $n$ samples is used, usually ($n = 60$) consisting of four circles. Then, a Gaussian smoothing is performed with standard deviation equal to the distance between the points of the same circle (Fig. 2.6). In addition, the local gradient over the $k$ is calculated and the estimation of the direction of $k$ is obtained. This mechanism gives to the method, rotation invariance. The description of point $k$ is formed by the bit vector $d_k$ which is assembled from the short-distance pairwise intensity comparisons. Finally, to compare two BRISK descriptors, the Hamming distance

is employed. [22]

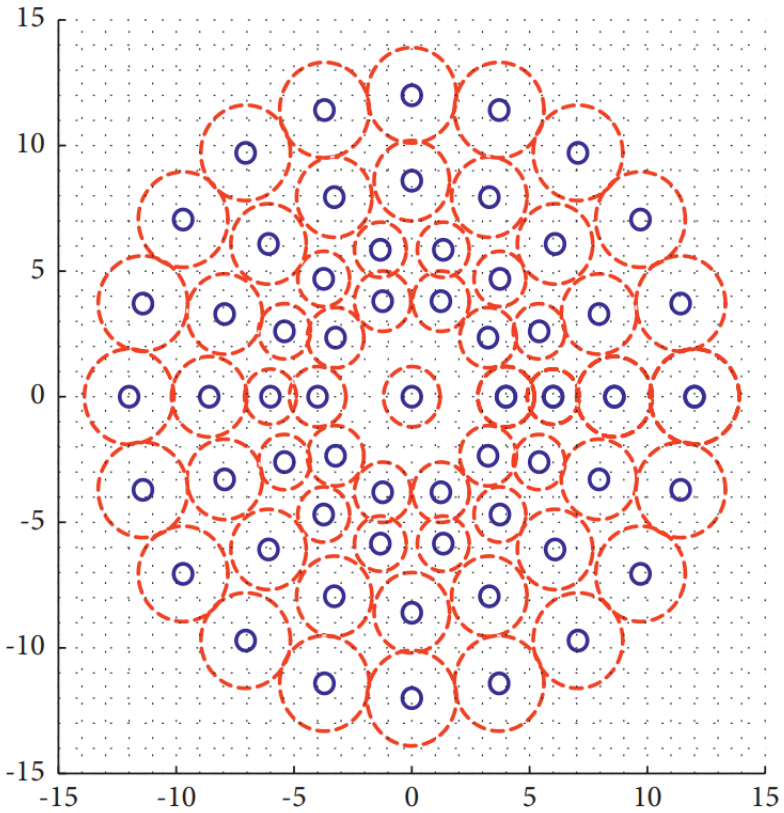


Figure 2.6: The figure represents the BRISK description mechanism. Sampling pattern with $n = 60$ points is shown; the small blue circles denote the sampling locations and the red dashed circles are the Gaussian kernels that are used to smooth the neighbourhood. [54].

Fast Retina Key-point (FREAK) is a binary feature descriptor, suggested by Alahi et al. in 2012 [55]. It is inspired by the human visual system, or more precisely, by the retina. The binary descriptor is produced with the retinal sampling grid that is a circular pattern which has a higher density of points near the center. Indeed A human vision-like search (called a saccadic search) is used to select relevant features. Moreover FREAK compensate rotation changes measuring the orientation in a similar way to BRISK.

Alcantarilla et al. illustrated the KAZE algorithm [56] which consists of detecting and describing 2D features in a non-linear scale space to obtain better distinction and location precision. This detection method is based on the calculation of the Hessian matrix and the description is made computing the key point orientation. KAZE descriptor is invariant to scale and rotation, but it is computationally expensive. For this reason AKAZE (Accelerated-KAZE) was implemented by the same authors [57] to improve the previous

method. As the name indicates it is an accelerated version of KAZE, providing similar or even better performances in some scenarios.

In the last decade, several reviews have been conducted. One of the most relevant work was published by Mukherjee et al. in 2015 [22]. The authors showed a comparison between some combinations of detectors and descriptors in a practical scenario. The analysis is based on the algorithm invariance against image transformations such as illumination changes, blurring, rotation, scaling, viewpoint changes, exposure, combined scaling and rotation, and combined viewpoint changes. The considered architectures are the following: BRISK, CENSURE+BRIEF, FAST+BRIEF, MSER+BRIEF, ORB, SIFT and SURF. The comparison was made in terms of average positional error and computational time for motion recovery. It shows that ORB was characterized by the lower computational cost but the worst accuracy in scaling as opposed to SIFT.

In 2018 Tareen et al. [18] evaluated accurately the performance of several feature detectors and descriptors applied in the image registration scenario. This article presents an exhaustive comparison of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. The experimental results provide rich information and various new insights that are valuable for making critical decisions in vision based applications. The comparison is provided on different levels. SIFT, SURF, and BRISK are found to be the most scale invariant feature detectors that have survived wide spread scale variations. ORB is found to be the least scale invariant. ORB, BRISK, and AKAZE are more rotation invariant than others. ORB and BRISK are generally more invariant to affine changes as compared to others. SIFT, KAZE, AKAZE, and BRISK have higher accuracy for image rotations as compared to the rest.

Win and Kitjaidure [58] applied the image stitching methodology based on features detection, description and matching in the biomedical field. This feature-based system is used to stitch high-resolution biomedical images with a short processing time. The process is divided five stages: pre-processing, feature extraction, feature matching, homographic estimation, and image stitching. In the feature detection stage, a method based on ORB features is considered and compared to many different feature detectors, such as Harris Corner Detector, SIFT, and SURF technology.

## 2.2. Keypoints Matching

Correspondences between points in images are essential for estimating the 3D structure and camera poses in geometric computer vision tasks such as Simultaneous Localization and Mapping (SLAM) and Structure-from-Motion (SfM). Such correspondences are gen-

erally estimated by matching local features, a process known as data association. [19]

Three feature point matching algorithms commonly used in literature are: Brute Force matching algorithm (BF), K-Nearest Neighbors (kNN) and Fast Library for Approximate Nearest Neighbors (FLANN). Their performances vary basing on the feature detector and descriptor with which they are coupled.

The matching process of BF consists of taking the descriptor of one feature in the first set (referred to $x_m$) and matching it with all other features in the second set (referred to $x_r$). The correspondence is done calculating the distance between the points, finally the closest key point is returned and the match is completed. This is computed for all the keypoints detected in the two images. Different types of distance can be implemented for this purpose. The metric distance used by default is the Euclidean distance (Eq. 2.2) between two vectors $M$ and $R$, where $M$ is the set of key points of the frame $x_m$ and $R$ is the set of key points of the frame $x_m$. Another metric that is often involved to perform the BF matching is the Hamming distance. The Hamming distance (Eq. 2.3) Hamm($M$, $R$) between the features vectors $M$ and $R$, both of length $n$, is the number of positions $u$ such that $M(u)$ is different from $R(u)$ and is a value from 0 to $n$. [59, 60, 60].

$$dist(M,R) = \sqrt{\sum_{i=o}^{n}(M_i - R_i)^2} \tag{2.2}$$

$$Hamm(M,R) = \sum_{i=0}^{n}(M_i == R_i)?1:0 \tag{2.3}$$

KNN method follows the same principle of BF matcher. It considers the keypoints descriptors of $x_m$ and $x_r$ and it searches for the k best matches in the nearest neighbour set of pixels around the points of interest. This is employed through a k-d tree, which is a binary multilevel tree where each note represents a subfile of the query descriptors. Then, the Euclidean distance (Eq. 2.2) between the descriptor vectors is calculated. Finally, a good match is detected if it meets the condition proposed by Lowe, expressed in equation below. [59]

$$Lowe's\ Condition = \begin{cases} if\ dist(M,R) \leq M - 0.7R & Good\ Match \\ if\ dist(M,R) > M - 0.7R & Reject\ Match \end{cases} \tag{2.4}$$

Where the value 0.7 is a threshold selected by the user.

To resume, respect to the BF matching algorithm, which returns the best match, KNN gives the k best matches, where k is indicated by the user. [61]

The FLANN feature matcher includes a set of algorithms for fast nearest neighbour search in large data sets and high-dimensional features. In the process of feature point matching, the procedure of FLANN is using the selected algorithm to calculate and find the specific feature point with the nearest distance of descriptors as the matched feature point.

In conclusion the BF and the KNN algorithm work well for small sets of key points, but they tend to be slower when the quantity of features increases. However, FLANN shows a lower computational cost with a large set of keypoints, respect to the other two methods. [59]

For example Noble et al. delineated a comparison of OpenCV's feature detectors and feature matchers. [62] SIFT, SURF, BRISK, ORB, KAZE, and AKAZE feature detectors, coupled with BF and FLANN feature matchers, have been implemented. The evaluation of their performances was based on the number of features detected and detection time related to the number of features matched and the matching time. In conclusion they noticed that BF is the algorithm that matches more features and the combination of BRISK and BF provided a reasonable balance between the number of features detected and matched and a relatively short computation time.

Among all these algorithms, the commonly used is SIFT because of its stability to rotation and scale. However it is not possible to establish the best method or an ideal one for specific purposes also because each feature detector presents some issues. For example SIFT is characterized by scale and rotational invariance but is associated to an higher computational cost. On the contrary ORB is faster in key points detection but less accurate.

keypoint detection, description and matching phase are connected because the final result is highly dependent on their coupling. Moreover the use of a feature detector reduces the search space of matching, for this reason it could be better to implement a unique architecture which faces to this problem. [63]

In addition a traditional feature detector may fail to extract enough keypoints in specific settings characterized by various factors such as poor texture, repetitive patterns, viewpoint change, illumination variation, and motion blur. These problems make also the matching phase particularly challenging. [19] These conditions are especially prominent in neurosurgical environments, where low-texture regions or repetitive patterns sometimes

occupy most areas in the field of view. Without repeatable keypoints, it is impossible to find correct correspondences even with perfect descriptors. [63]

## 2.3. Learning Based Methods

Deep learning (DL) has achieved rapid developments in computer vision and image processing such as object detection, image identification and image classification. DL can be used in the method of image registration that is classified into the intensity-based method and the feature-based method. In the classical deep neural intensity-based method, a general solution is that the deep learning is used as an iterator to optimize the loss function between the reference image and the floating image to estimate the transformation function. When the loss value reaches the required range, the transformation matrix is obtained. [63]

To improve the invariance to deformation, semi- and selfsupervised learning is also attempted using GANs and autoencoder. However, intensity-based methods are unsuitable for large displacement problems which are handled by feature-based methods.[63]

Learning schemes have been used in feature-based image registration to detect keypoints, describe keypoints, and to estimate transformation between images. The FAST detector firstly uses machine learning techniques to classify a pixel point into a corner point or not using decision trees. [64] For FAST, learning and supervised transformation have been adopted for the high repeatability and to speed up the convergence but the principal drawback is the learning algorithm high dependency on the training data, which could not cover all possible corners. [63]

Another example of DL network applied for keypoints detection is the Temporally Invariant Learned Detector (TILDE), proposed by Veredie et al. [65]. It designed to detect repeatable keypoints in images with drastic illumination changes, conosidering images captured in different conditions, such as different moments of the day, different weather or seasons. These images constituted the training set. Then SIFT was used to detect and locate the position of keypoints. A linear regressor is trained to predict a score map, whose value is greater than a threshold, a keypoint is identified. Despite the promising architecture, TILDE only remains a state-of-the-art approach for keypoints detection in the presence of illumination changes, as it is not characterized but scale and rotation invariance. [63]

Learn Invariant Feature Transform (LIFT) was attempted by Yi et al. [66] It is characterized by a learning detection, orientation estimation, and keypoints description in a

unified pipeline that consists of three convolutional neural networks (CNNs).The authors claimed that LIFT can be regarded as a trainable SIFT. The training procedure concerns the descriptor first, then the orientation estimator for the descriptor, and finally the detector. LIFT is an improvement of TILDE that is learned to robustly detect features in spite of illumination changes. However, the learning only carries on a dataset without viewpoint and scale changes, in this way the method does not learn the scale invariance of the detector in the training process. Although the method proposed an effective strategy to train each component individually, resulting in running jointly, the further objection is to look into performing the method over the whole image instead of pre-extracted patches. [63]

DeTone et al. designed the SuperPoint architecture, inspired by recent advances in applying DL to keypoint detection and descriptor learning. [67] They described a fully-convolutional neural network (FCNN) architecture for keypoint detection and description trained using a self-supervised domain adaptation framework called Homographic Adaptation. Their aim was to compute both keypoints and descriptors in a single network in real-time. A similar approach to Homographic Adaptation was presented by Honari et al. [68] under the name "equivariant landmark transform." Also, Geometric Matching Networks [69] and Deep Image Homography Estimation [70] use a similar self-supervision strategy to create training data for estimating global transformations. However, these methods lack in keypoints correspondences. Moreover different from LIFT, this method performs on full-sized images to computer interest point at pixel level and associated descriptors in one forward pass instead of relying on preextracted patches. [63]

The principal analysed methods for keypoints detection and description are summarised with their fundamental proprieties (invariance to scale, rotation and illumination) in Table 2.1.

In this scenario, with the developments of DL, Ma et al. [71, 72] have reviewed and proved that Convolutional Neural Networks (CNNs) are the mostly used deep net architectures in keypoints detection, description, and matching in comparison with other models. The principle of the deep learning-based detector is to construct a response map and then search keypoints in it. Moreover the detector is trained in a differentiable way and under the geometric transformation constraints between images. This type of method can be classified into supervised, self-supervised, or unsupervised methods. [63]

As Abbadi et al. asserted in [73], image stitching and registration is still an important challenge in many image processing and computer vision tasks. Unfortunately, there isn't still an appropriate algorithm that provides a precise and accurate panoramic image. In-

Table 2.1: This table resumes the principal characteristics of the most relevant keypoints detectors and descriptors presented in literature, both traditional and learning based. The reported features detectors and descriptors are: BRISK, ORB, KAZE, SIFT, SURF, LIFT and Superpoint. The keypoints detectors are: Harris Corner Detector, FAST and TILDE. The keypoints descriptors are: BRIEF and FREAK. The analysed proprieties are: scale, rotation and illumination invariance both for detection and description algorithms. [54]

| **Detector** | Scale Invariance | Rotation Invariance | Illumination Invariance | **Descriptor** | Scale Invariance | Rotation Invariance | Illumination Invariance |
|---|---|---|---|---|---|---|---|
| Harris | × | ✓ | ✓ | - | - | - | - |
| FAST | × | × | ✓ | - | - | - | - |
| BRISK | ✓ | × | ✓ | BRISK | ✓ | ✓ | ✓ |
| ORB | × | ✓ | ✓ | ORB | × | ✓ | ✓ |
| - | - | - | - | BRIEF | × | × | ✓ |
| - | - | - | - | FREAK | × | × | ✓ |
| KAZE | ✓ | ✓ | ✓ | KAZE | ✓ | ✓ | ✓ |
| SIFT | ✓ | ✓ | ✓ | SIFT | ✓ | ✓ | ✓ |
| SURF | ✓ | ✓ | ✓ | SURF | ✓ | ✓ | ✓ |
| TILDE | × | × | ✓ | - | - | - | - |
| LIFT | × | ✓ | ✓ | LIFT | × | ✓ | ✓ |
| SuperPoint | ✓ | ✓ | ✓ | SuperPoint | ✓ | ✓ | ✓ |

deed the obtained mosaic depends strongly on the characteristics of the available dataset. In particular the typical conditions of a surgical setting such as viewpoints change, illumination variation and motion blur, as I mentioned before, are factors likely to be responsible for inaccuracies in the reconstruction process. However, the technology is advancing with the aim of achieving high accuracy results.

In the biomedical scenario Bano et al. [74] proposed a deep learning-based mosaicking framework for diverse fetoscopic videos captured from different settings such as simulation, phantoms, ex vivo, and in vivo environments. The idea is that fetoscopic mosaicking can help in creating an image with the expanded field of view which could facilitate the clinicians during the twin-to-twin transfusion syndrome (TTTS) treatment. Placental panorama is built starting from segmented vessels; but the highly dependence on the correctness of segmentation, makes this approach problematic and not robust enough.

This thesis aims to apply the mosaicking technique to a neurosurgical setting to deal with the presented issue of low visibility in neurosurgery environment.

The contribution of this thesis work can be summarized as follows:

1. To the best of our knowledge, it represents the first application of mosaicking on a neurosurgical dataset.
2. A robust self-supervised method for keypoints detection and description to be compared with the traditional algorithms presented in literature.
3. An attentional graph neural network based on [19], trained in a self-supervised way on intra-operative images for keypoints matching.

4. The management of unexpected situations in the surgical room (like accidentally bump the microscope). These situation cause fast movements of the camera that could damage generating mosaic. An homography check and filter is performed in order to save the reconstruction, obtained until the unexpected event occurs.

# 3 | Materials and Methods

The purpose of video mosaicking is to combine consecutive frames of a video sequence, in which each frame shows only an partial local view of the field of interest. It allows to obtain a broader view of the same scene. [17]

The classical mosaicking approach is characterized by the following four stages: i) Feature detection and description; ii) Feature matching with outlier rejection; iii) Homography estimation; and iv) Image warping and blending. In this process each step is essential for the correct execution of the next ones. [18]

A features detector is an algorithm that searches for keypoints into an image. Keypoints can range from a single pixel to edges, corners, contours, blobs, junctions and lines; they are expressed by a system of coordinates and represent the most significant pixels of the selected image. [18]

Once the keypoints have been extracted from the image, the feature description phase is applied. A descriptor is a vector, needed to assign to each key point a distinctive identity which allows its effective recognition for matching. The features description is based on unique patterns possessed by the neighboring pixels of each keypoint. [19]

Given a frames sequence, each images pair is considered. The frame pair consists of a moving image ($B$) to be registered respect to the reference image ($A$).

Keypoints detection and description are computed for each image of the pair. The next stage of the mosaicking process is the feature matching between the keypoints of the two images. It aims to establish a correct correspondences from the keypoints sets. [20]

Afterwards, RANdom Sample Consensus (RANSAC) algorithm is employed for the homography estimation. The homography matrix is applied to the moving image $B$ and it is used to merge $A$ with $B$. This step is called image warping. [18]

After this overview of the mosaikcing process, in the next sections the proposed method is described. It is an end-to-end learning-based architecture, used to perform keypoints detection description and matching, and it is called NeuroGlue.

## 3.1. Keypoints Detection and Description

In NeuroGlue, the keypoints detection and descriptor phase are combined in a fully-convolutional neural network (FCNN), which is called SuperPoint. [67]

In particular this network is able to detect robust and repeatable keypoints and to attach a fixed dimensional descriptor vector to each keypoint for further processing, such as image matching.

The first step that is applied is the dimensionality reduction of each input image, which is performed with a single shared-encoder. The encoder consists of convolutional layers, spatial downsampling via pooling and non-linear activation functions. After the encoder, the architecture splits into two decoders: one for keypoint detection and the other for keypoint description. Both decoders operate on a shared and spatially reduced representation of the input performed in a VGG-style.

In this way most of the network's parameters are shared between the two decoders, and this represents an evolution respect to the traditional systems. Indeed the classical methods first detect keypoints, then compute descriptors and lack the ability to share computation and representation across the two tasks. [67]

For the image pair considered, the keypoints $p^A$ and the relative descriptors $(d)^A$ for the frame $A$ and $p^B$ with the descriptors $d^B$ for the frame $B$ are obtained. Fig. 3.1.

## 3.2. Graph Based Matching

NeuroGlue is based on an Attentional Graph Neural Network (AGNN) that concerns the matching computation, and it takes inspiration from SuperGlue network. [19]

In the first step keypoints and descriptors obtained from the SuperPoint network, are subsequently combined into a single vector using a keypoints encoder. In this way a first features representation is achieved ($f_i^A$ for image A and $f_i^B$ for image B).

After the positional encoding, coarse features are the input of a module composed by a Self Attentional layer and a Cross-Attentional layer. In neural networks, attention is a technique that mimics cognitive mechanisms, using an encoder-decoder architecture. The rationale of attention mechanism is to discriminate which input points are important by assigning them a high weight factor, and to diminish less relevant point. In Self-Attention layers the transformer concentrates on mapping meaningful points inside the same image, while in Cross-Attention layer relevant points between the two input images are mapped

ad enhanced. The output of the described module facilitates matching, due to the high feature dependency on position and context. In particular the alternation of the two Attentional Aggregration layers allows to develop the keypoints connection, making it stronger and stable for matching. (Fig. 3.1) From this process, matching descriptors (indicated with $m_i^A$ and $m_i^B$) are achieved.

The affinity between the correspondeces is defined with a score matrix ($S_{ij}$), which is built making the dot product between the matching descriptors of each image, previously obtained.

The score matrix is used also to filter out the invalid matches and relative keypoints, present due to occlusions and noise. This procedure is carried out with the introduction of a dustbin, as indicated in Fig. 3.1.

Finally, the Sinkhorn algorithm is applied as an optimization layer in order to increase the reliability of the optimal transport estimation for matching computation. Indeed optimal transport tool is used to find the minimal cost in probability distribution data pairs. Sinkhorn algorithm is an iterative process that normalizes the score matrix around the rows and the columns. From the obtained result, the matches can be extracted.[19]

## 3.3. Homography Estimation

After that the valid matches and the correspondent keypoints are extracted, RANdom Sample Consensus (RANSAC) and Levenberg-Marquardt (LM) algorithm are jointly applied to estimate the homography. The objective of RANSAC is to select the optimal set of keypoints and filter out outliers that do not fit with a defined type of transformation, and this procedure is based on the back projection error minimization.

The samples which satisfy the model are called inliers and the corresponding set constitutes the consensus set. The samples which do not satisfy the model are called outliers, that are rejected. The model which produces the larger consensus set is considered as valid and the corresponding inliers are kept for further processing. [21]

Homography is then estimated using the openCV function $cv2.estimate2DAffine()$ or $cv2.findHomography()$. The homography $H$ is a $3 \times 3$ matrix, which provides the relative transformation of $B$ respect to $A$. [18, 19]

The two openCV functions receive the same inputs: the array of coordinates of the keypoints extracted from image $A$ ($p^A_{i_filtered}$), the keypoints coordinates detected in image $B$ ($p^B_{i_filtered}$), the function $cv2.RANSAC$ to correctly fit the data into the model and the

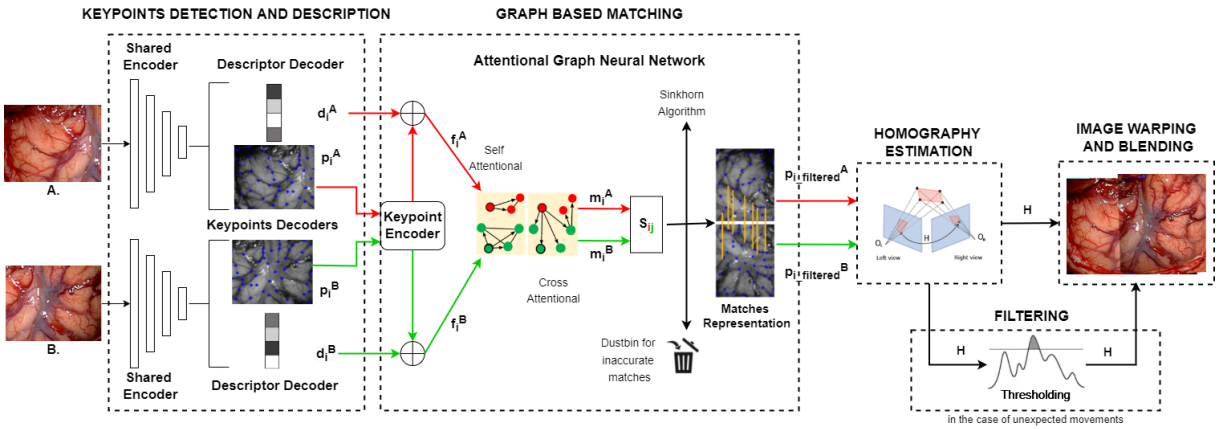


Figure 3.1: Overview of the proposed framework for neurosurgery mosaicking, as described in Chapter 3. The first block concerns keypoints detection and description phases. Each keypoint and descriptor of the image $A$ are indicated with $p_i^A$ and $d_i^A$ respectively. The same idea is applied to the image $B$ (Sec. 3.1). These outputs are then combined using a keypoint encoder in order to obtain a unique feature vector for each image ($f_i^A$ and $f_i^B$). Their combination is indicated with $\bigoplus$. $m_i^A$ and $m_i^B$ are the matching descriptors obtained from the alternation of Self and Cross Attentional Layers. The affinity between the correspondences is represented by the score matrix $S_{ij}$, which is also used to filter out invalid matching with a dustbin. Matching optimization is performed with the Sinkhorn Algorithm. (Sec. 3.2) Removing the key points relative to invalid matches, $p^A_{ifiltered}$ and $p^B_{ifiltered}$ are identified and are employed for the homography estimation ($H$)(Sec. 3.3), essential for image warping and blending (Sec. 3.4). Also the optional filtering stage is represented. It is applied for the management of unexpected movements of the camera as it is described in Sec 3.5.

$ransacReprojThreshold$, back projection error threshold imposed equal to 2.

The output of $cv2.estimate2DAffine()$ is a $2 \times 3$ matrix, which is combined to the last row of the identity matrix to create the $3 \times 3$ homography matrix $H$. Instead the output of $cv2.findHomography()$ is directly $H$. Sperimentally it was observed that the best results are obtained with $cv2.estimate2DAffine()$ function.

In general the homography matrix represents the transformation between the points of two different planes as indicated in the following equation:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3.1}$$

In the reported equation 3.1 $H$ represent between the transformation of $P'(x', y', z')$ respect to the reference plane $P(x, y, z)$. $H$ presents eight degrees of freedom because it is generally normalized with:

$$h_{33} = 1 \tag{3.2}$$

or

$$h_{11}^2 + h_{12}^2 + h_{13}^2 + h_{21}^2 + h_{22}^2 + h_{23}^2 + h_{31}^2 + h_{32}^2 + h_{33} = 1 \tag{3.3}$$

The homography is created combining different components: the camera intrinsic matrix (which depends on the focal length); rotation matrices around the X,Y,Z axis and the translation array in X,Y,Z directions. $H$ is essential for image warping which is the following step of the process.

## 3.4. Image Warping and Blending

Image warping and blending is performed combining $B$ with $A$, basing on the matrix $H$ previously computed, until all matched feature points are aligned. It is performed using the $cv2.warpPerspective()$ openCV function. This function receives as input parameters, the considered frame ($B$), $H$, and the dimension of the canvas image, in order to attach the image $B$ accordingly transformed basing on $H$ in the correct position of the canvas respect to $A$, already located on it. A canvas is a black image, characterized by greater dimensions respect to the input frames, and it is used to host the generating panorama (Fig. 3.1).

Considering a frames sequence, for example extracted from a a video, each frame pair is considered and the same steps, described above, are applied to them.

In particular $cv2.warpPerspective()$ function needs as input the absolute homography matrix, which is obtained performing the product between all the relatives homographyes ($H$) which characterize each image pair of the sequence.

In this way each selected frame of the sequence is transformed basing on absolute homography matrix and then attached to the canvas in the correct position. The canvas with the attached frame becomes the new canvas for the next frames, and this process continues until the end of the frames sequence, obtaining a sort of map composed of smaller images, called mosaic or stitched image. [22]

## 3.5. Filtering

Neurosurgery procedures are characterized by limited movements of the microscope. Indeed broad and rapid movements are unlikely because surgeons used to work with high magnifications in a reduced operative field. However, it could happen to mistakenly impact the microscope generating very fast movements that could make the registration algorithm to fail. Any abnormal movements of the camera, both translating and rotational, could generate distortions and reconstruction errors. For this reason a filtering stage is implemented.

After $H$ estimation, the Singular Value Decomposition (SVD) is performed. SVD procedure is based on the matrix factorization using eigenvalues and eigenvectors.

Experimentally it was demonstrated that if an unexpected movement occur one or more decomposed values show a steep increase characterizing an abnormal homography.

In particular with SVD, six parameters are obtained: $t_x$ and $t_y$ which reflect the translating movements of the camera, $s_x$ and $s_y$, related to the scaling and $\gamma$ and $\theta$ that translate the rotational transformations.

These parameters are computed as follows, considering the homography $H$, as indicated in Eq. 3.1.

$$E = \frac{(h_{00} + h_{11})}{2} \tag{3.4}$$

$$G = \frac{(h_{10} + h_{01})}{2} \tag{3.5}$$

$$H_{val} = \frac{(h_{10} - h_{01})}{2} \tag{3.6}$$

$$Q = \sqrt{E^2 + H_{val}^2} \tag{3.7}$$

$$R = \sqrt{E^2 + G^2} \tag{3.8}$$

$$a_1 = \arctan(\frac{G}{E}) \tag{3.9}$$

$$a_2 = \arctan(\frac{H_{val}}{E}) \tag{3.10}$$

$$s_x = Q + R \tag{3.11}$$

$$s_y = Q - R \tag{3.12}$$

$$\theta = \frac{a_2 - a_1}{2} \tag{3.13}$$

$$\gamma = \frac{a_2 + a_1}{2} \tag{3.14}$$

$$t_x = h_{02} \tag{3.15}$$

$$t_y = h_{12} \tag{3.16}$$

The correlation among the different parameters is assessed by computing the Pearson's Correlation ($\rho$) reported in Table 3.1.

Table 3.1: Pearson correlation ($\rho$) among parameters obtained through SVD of the homography transformation computed in Sec. **??**

| $\rho$ | $t_x$ | $t_y$ | $s_x$ | $s_y$ | $\gamma$ | $\theta$ |
|---|---|---|---|---|---|---|
| $t_x$ | 1 | 0.999 | 0.999 | -0.999 | 0.182 | 0.997 |
| $t_y$ | 0.999 | 1 | 0.999 | -0.999 | 0.169 | 0.996 |
| $s_x$ | 0.999 | 0.999 | 1 | -0.999 | 0.190 | 0.996 |
| $s_y$ | -0.999 | -0.999 | -0.999 | 1 | -0.151 | -0.995 |
| $\gamma$ | 0.182 | 0.169 | 0.190 | -0.151 | 1 | 0.223 |
| $\theta$ | 0.997 | 0.996 | 0.996 | -0.995 | 0.223 | 1 |

Two parameters are correlated when the value of $\rho$ is close to $\pm 1$. For this reason $\gamma$ is selected. To achieve a more complete and robust analysis also $t_x$ is considered.

At each iteration, $t_x$ and $\gamma$ are compared with two thresholds, respectively, which are experimentally selected. An abnormal change in the homography (one or both values are over-threshold) interrupts the registration procedure, discarding the associated frame since a new valid frame is present. Therefore this homography check allows to restore the mosaic, obtained before the unexpected event occurs.

## 3.6. Dataset

The dataset available was supplied by Humanitas Research Hospital of Milano, in which several videos was captured from a Carl Zeiss Surgical GmbH microscope. A physical


Figure 3.2: This image shows the characteristics of the Carl Zeiss Surgical GmbH microscope. It was taken form the ZEISS company website. [75]

representation of the SM is illustrated in Fig. 3.2.

The extracted frames have original dimension of $720 \times 576$.

In particular three videos are available: Video1 contains 1677 frames, Video2 is composed by 667 frames (it is the shorter one) and Video3 has 3543 frames. Some frames belonging to these videos are reported as examples in Fig. 3.3.

Images captured from a surgical setting are characterized by regular patterns (for the blood vessels structure), viewpoint changes, illumination variations, and motion blur. Moreover, due to the intrinsic characteristics of neurosurgical environment images of this setting are dishomogeneous in illumination. In general the operative room is widely illuminated but the light introduced by the SM hits specific areas of the brain, making them brighter than the other areas of the surgical field. This characteristic is visible in Fig. 3.3. Thus, when images are stitched together, this imbalance in the illumination affects the reconstruction.

In general all the presented factors make the classical mosaicking methods not enough robust and stable, and the reconstruction process challenging.

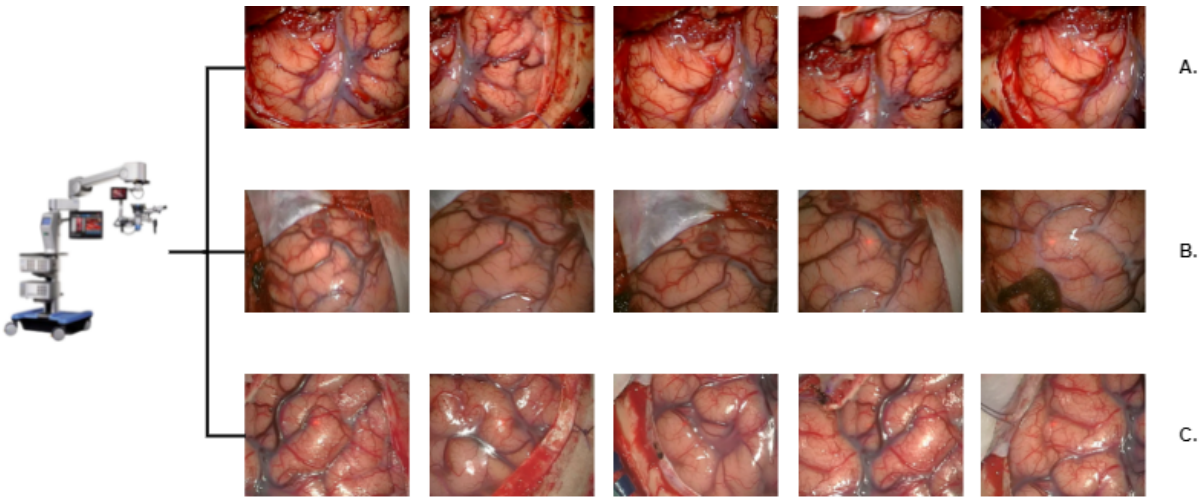


Figure 3.3: This figure shows some frames belonging to the three extracted videos: A)Video1, B)Video2 and C)Video3.

# 4 | Experimental Protocol

## 4.1. Training Phase

The NeuroGlue training is performed on a dataset of 6144 non-overlapped patches with dimension 256x256, extracted from Carl Zeiss Surgical GmbH microscope videos.

NeuroGlue is trained end to end in a supervised way [19], for 300 epochs.

The Keypoints detection and Description network training is based on a method called Homography Adaptation. [67] It consists on the random homographies generation that are used to warp copies of the input image and combine the results. A random homography is generated combining less expressive and simpler transformations, as Fig. **??** shows.

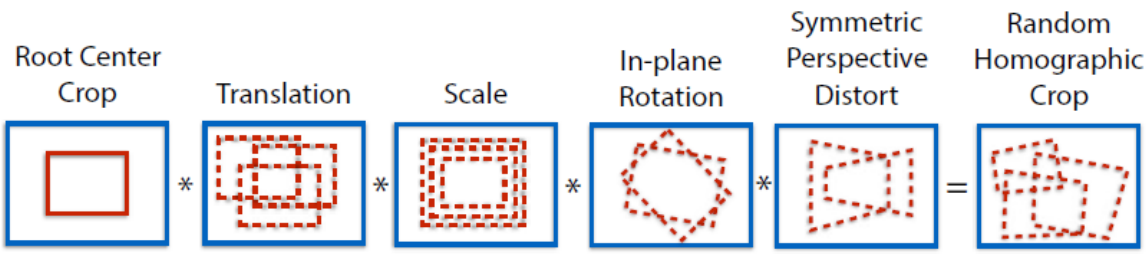


Figure 4.1: This figure illustrates how to generate a random homography. The procedure includes the composition of simpler transformations, such as rotations, scaling and symmetry perspective distortions. These simpler transformations are multiplied to obtain $M$.

The same image is deformed $L$ times, using $L$ different random homographies. In each warped image the keypoints are extracted, then the images are unwarped so an inverse transformation is applied to restore the input image and finally all the $L$ information obtained are merged to create the keypoints set.

The Attentional Graph Neural Network training is developed with the homography matrix computation and random image warping.

In particular, considering one frame $X$ selected randomly from the patches set, its key-

points and descriptors are computed with the previously trained SuperPoint architecture. [67]

A random warp transformation is applied to $X$: the image is deformed taking into account the expected movements of the microscope camera. It is done assigning to the patch corners new coordinates, generating random numbers between a range which is related to the patch dimension.

Experimentally it was shown that this procedure deforms the image simulating the microscope camera movements. In this way $X_{warped}$ is obtained.

Afterwards having $X$ and $X_{warped}$, the homography matrix $M$ which represents this transformation is computed. This is used to map the keypoints previously calculated with SuperPoint in $X_{warped}$. The keypoints in the border are then filtered out.

In this way the network learns to generate correct matches, since the keypoints correspondence is specially guaranteed, using the homography matrix $M$.

Fig. 4.2 represents two examples of matches generation during the training phase. The images show the computed correspondences between one random patch ($X$) and the same patch randomly distorted ($X_{warped}$). It is possible to observe the two reported examples of $X_{warped}$ ( Fig. 4.2.A.b and Fig. 4.2.B.b) are likely to resemble images obtained by tilting and shifting a video camera.

## 4.2. Ablation Study

To the best of our knowledge this work represents the first mosaicking application on a neurosurgical environment. For this reason, the analysis concerns the quality with which the methods presented in the literature for keypoints detection and matching (both traditional and learning based) fit a neurosurgical dataset.

The proposed method is compared with three traditional features detectors and descriptors: Binary Robust Invariant Scalable Keypoints (BRISK), Oriented FAST and Rotated BRIEF (ORB) and Scale Invariant Feature Transform (SIFT). These keypoints detectors are coupled with the K-Nearest Neighbors (kNN) matching algorithm. [18] These methods have been described in Chapter 2.

It was tested also the SuperGlue network coupled with SuperPoint, both pre-trained with the COCO dataset. [19, 67] This combination represents one of the most promising and accurate technique in the state of the art for learning-based keypoints detection and matching, outperforming others methods for the same purpose. [54]

It was tested also the SuperPoint network for keypoints detection, pre-trained with the COCO dataset [67] coupled with KNN algorithm for matching. SuperPoint architecture represents one of the most promising and accurate technique in the state of the art for learning-based keypoints detection. [54].

The compared methods are indicated as follows:

- Method 1: BRISK + KNN (it will be called BRISK)
- Method 2: ORB + KNN (it will be called ORB)
- Method 3: SIFT + KNN (it will be called SIFT)
- Method 4: SuperPoint + KNN (it will be called SuperPoint)
- Method 5: NeuroGlue (proposed architecture, based on Superpoint and Attentional Graph Based matching)

Experiment 4 and 5 are characterized by the same architecture for keypoints detection, but in experiment 5 the network is adapted to a neurosurgical domain (training performed with the surgical images). Moreover this comparison was done to underline the two different ways to compute matches: in Experiment 4 it is execute with the classical KNN algorithm, instead in Experiment 5 the precise coupling between the SuperPoint network and the Attentional Graph Based matching is shown.

## 4.3. Evaluation Metric

For the evaluation of the panorama reconstruction, obtained with the different methods, 5-frames Structural Similarity Measure (i.e. *SSIM*, indicated as $s$ in Equation 4.1) is computed. From the frame sequence $F$ of each video, one frame every five is selected. From this sampled list, consecutive pairs are taken. The second frame is transformed according to the relative transformation with the first. On the central 60% of the first frame and the second transformed frame, the metric is computed basing on the following equation:

$$s_{i \to i+n} = \text{sim}\left(w(\tilde{I}i, Hi \to i+n), \tilde{I}_{i+n}\right) \tag{4.1}$$

where $n$ is equal to 5, $w$ is the warping function, $sim$ is a similarity function, $I_i$ is the first image, $I_{i+n}$ is the second image, and $H$ is the relative transformation.

The reason of this 5-frames analysis is that two consecutive images would be too much

similar due to the small camera movements. The similarity between the two frames is expressed with a number between 0 and 1. A value close to 1 represents a high similarity between the two frames and so a less percentage of reconstruction errors.

However the use of this metric is limited. Let's consider one frame ($I_i$) to be compared with $5^{th}$ consecutive one ($I_{i+5}$). To each frame the relative homography $H$ is applied, so the transformed images are analysed. The final result obtained with a traditional mosaicking approach, such as BRISK or ORB, shows several deformations in the reconstructed mosaic. The presence of these errors is correlated to an homography $H$, characterized by abnormal parameters, obtained after the SVD process. Two frames with an irregular homography applied, result almost equal. In this way the *SSIM* computation doesn't reflect correctly the quality of the mosaic: since the two images are similar, the *SSIM* value is close to 1, but the obtained mosaic is very inaccurate. This situation is visually shown in Fig. 4.3.

To deal with this limitation a correction factor to $s$ is applied. In particular at each iteration the relative homography matrix $H$ is decomposed with the SVD procedure. The obtained parameters are analysed and compared with thresholds experimentally determined. If the SVD parameters are over-threshold the value of *SSIM* is drastically decrease to 0. The procedure is very similar to the one performed in the filtering stage, but the chosen thresholds are different. Moreover a check about the number of black pixels in the images is used to confirm the action.

*SSIM* is computed for the three traditional methods presented in literature (BRISK, ORB and SIFT), SuperPoint and NeuroGlue. This procedure is applied for the three videos available.

The Wilcoxon Signed-Rank test is performed in order to quantify from a statistical point of view the difference between values of *SSIM*, obtained with the different methods.

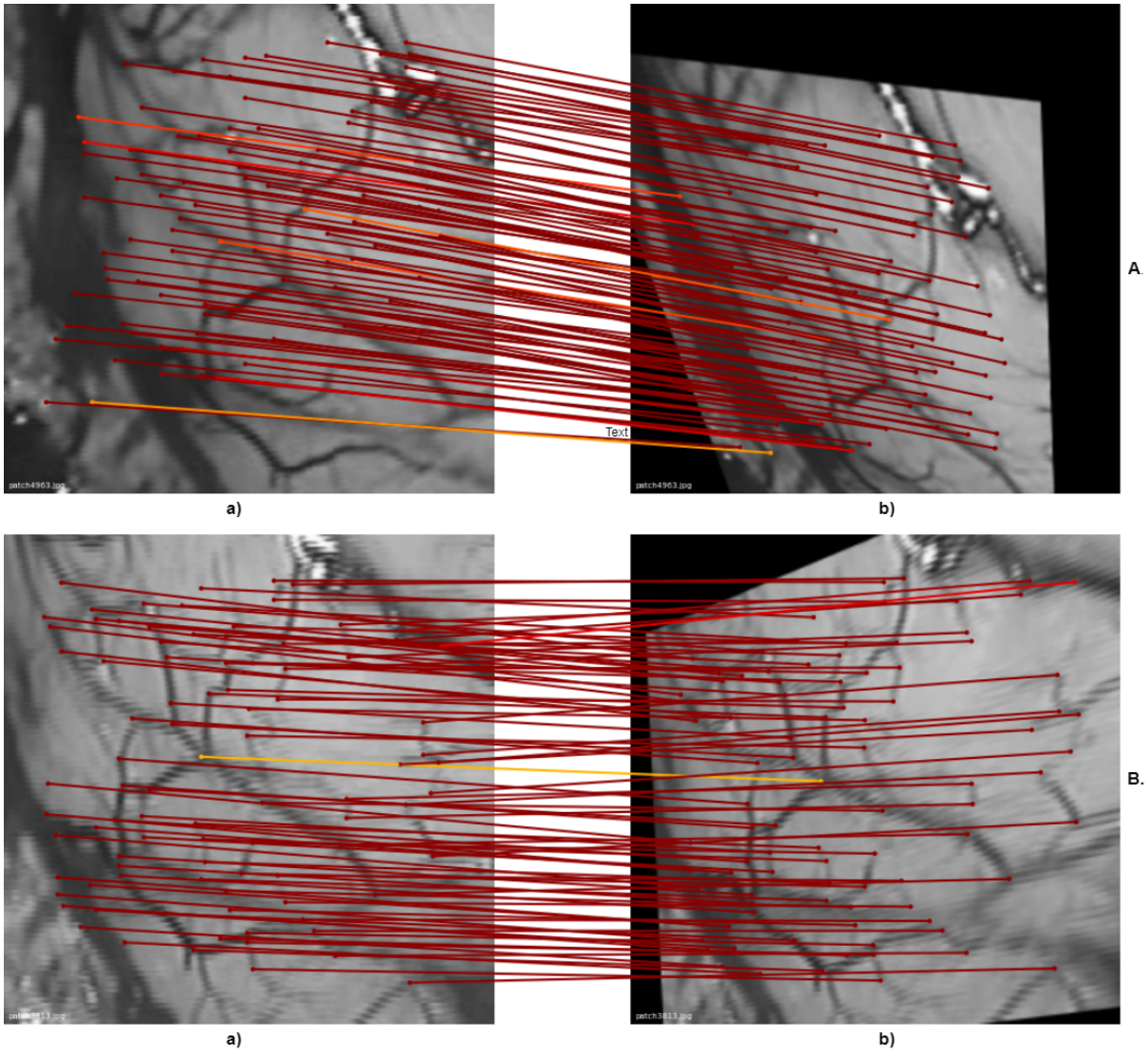


Figure 4.2: The image shows two examples of matches computation during training. The correspondences are obtained between one random patch ($A.a$) and its transformation ($A.b$). The same idea is applied for the example $B$.

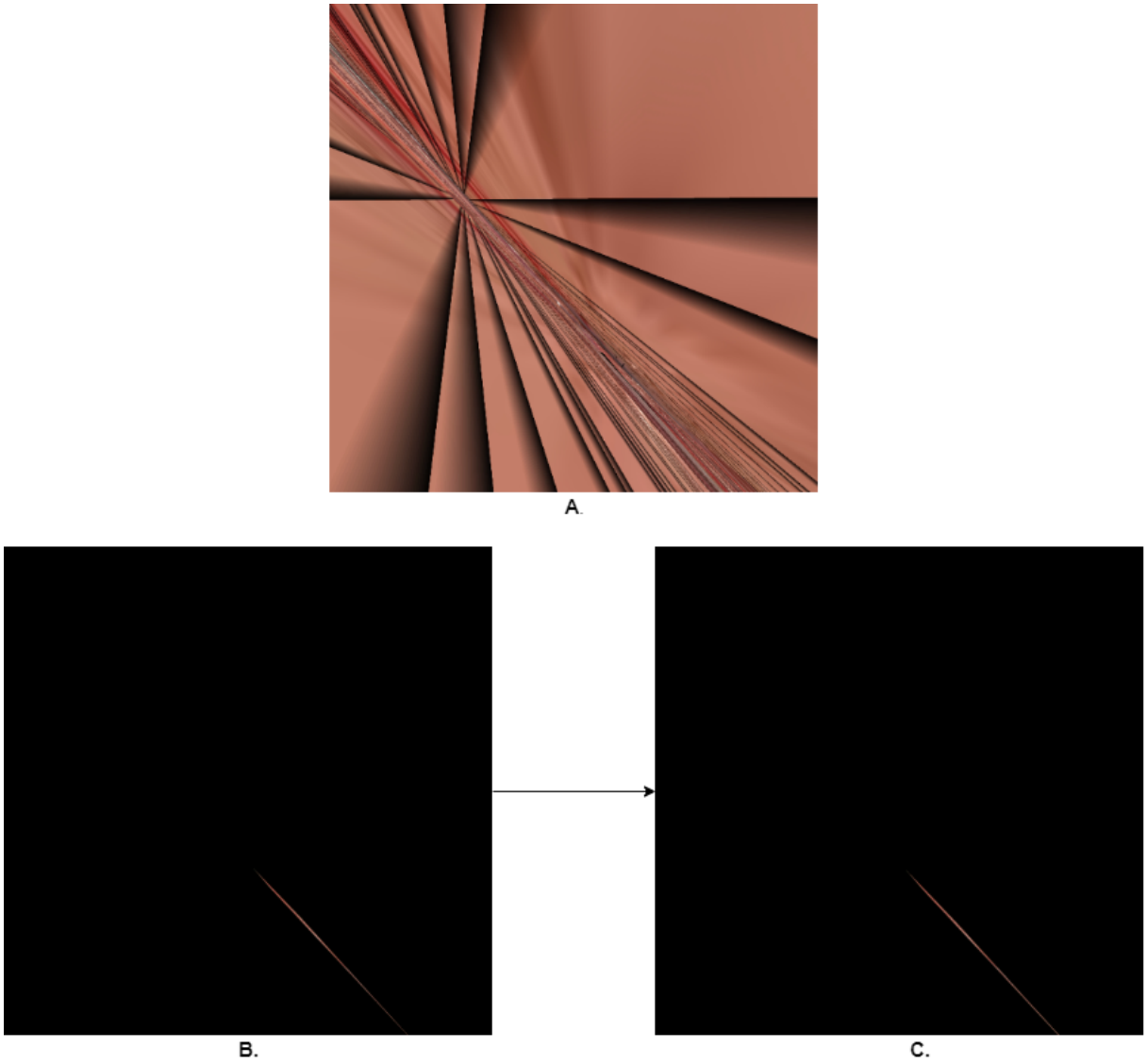


Figure 4.3: These images are related to Video3 dataset. A) is the ORB mosaic resulted at iteration number 2780. B) is the transformed frame number 2775 and C) transformed frame number 2780. The accuracy of the reconstruction is very low, but the two relative frames are almost equal. This is traduced to value of *SSIM* very close to 1 even if the mosaic is not adequate.

# 5 | Results

The obtained mosaics with the five different considered methods (BRISK, ORB, SIFT, SuperPoint, NeuroGlue) are reported in Fig. 5.1, for Video1, Fig. 5.2, for Video2 and Fig. 5.3, for Video3.

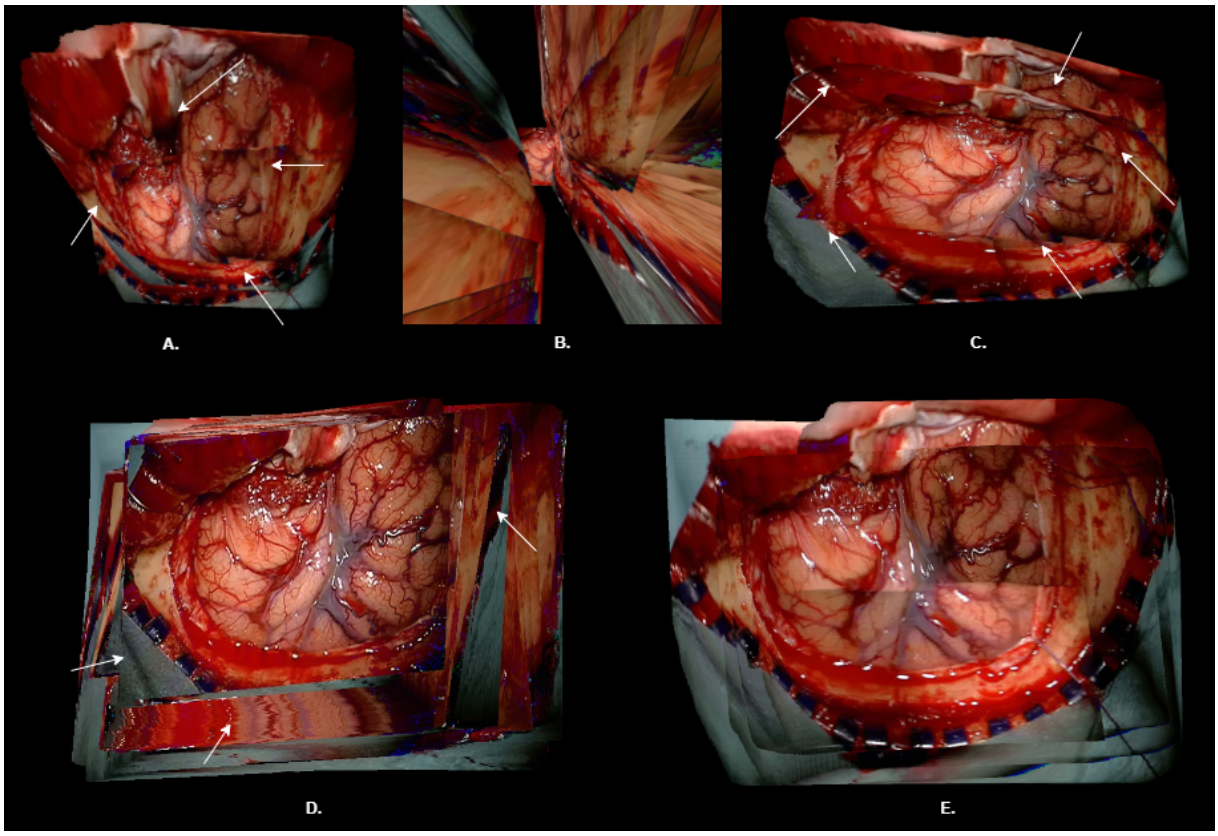


Figure 5.1: The figure shows the mosaics obtained with A.) BRISK, B.) ORB, C.) SIFT, D.) SuperPoint, E.) NeuroGlue. The white arrows indicates some inaccuracies and alignment errors in the registration. In this figure the results of video1 are reported.

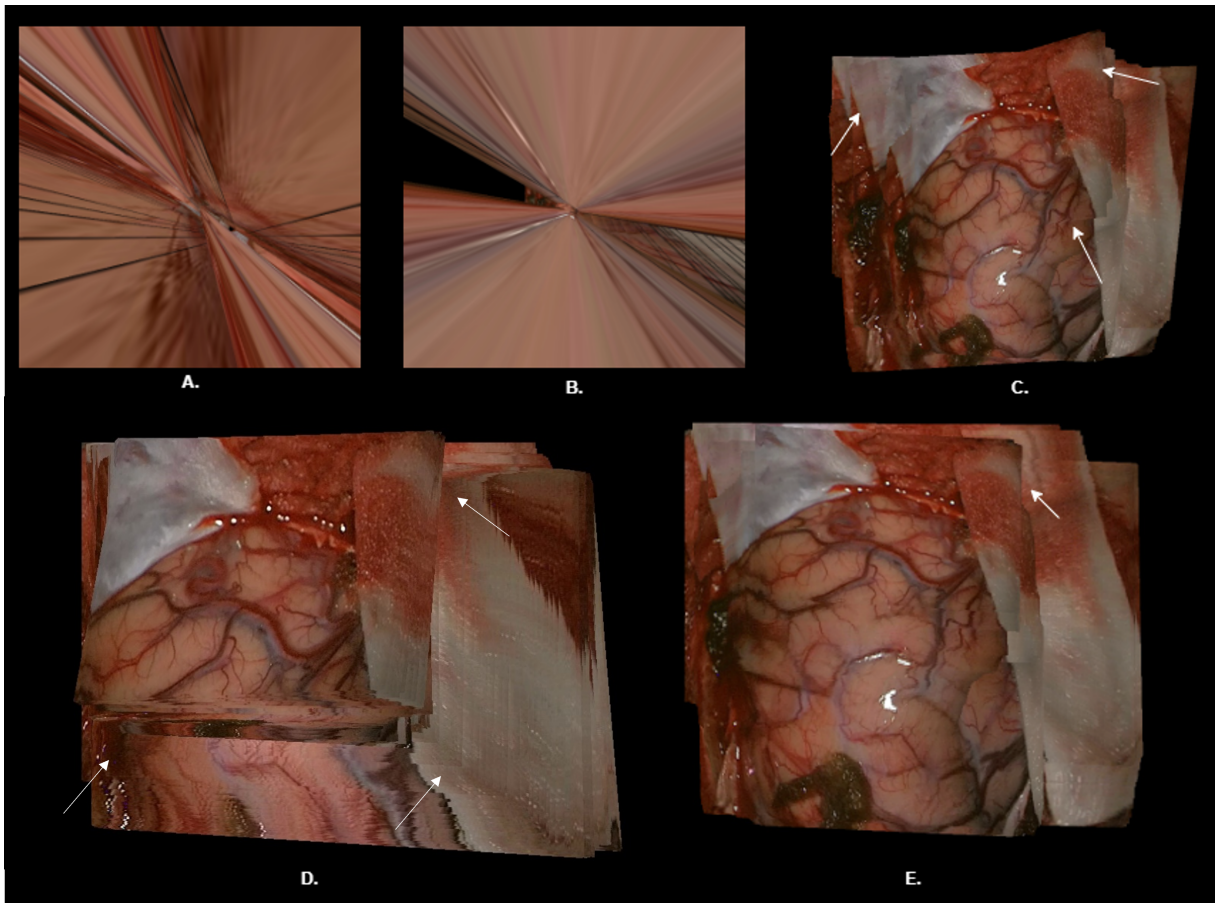


Figure 5.2: The figure shows the mosaics obtained with A.) BRISK, B.) ORB, C.) SIFT, D.) SuperPoint, E.) NeuroGlue. The white arrows indicates some inaccuracies and alignment errors in the registration. In this figure the results of video2 are reported.

The accuracy of the presented methods is quantified computing the Structural Similarity Metric (i.e. *SSIM*), as described in Sec 4.3. The mean and the average are computed for the structural similarity of each architecture; the results are reported in Table 5.1 for Video1, Table 5.2 for Video2 and Table 5.3 for Video3. With the obtained values the boxplots are build and represented in Fig. 5.4.

Fig. 5.5 and Fig. 5.6 show a visual example of the detected keypoints with BRISK, ORB, SIFT and NeuroGlue (using SuperPoint network). The trend of the extracted keypoints number is analysed respect to each frame of Video1 in Fig. 5.7.

Fig. 5.8 and Fig. 5.9 represent the relation between the number of computed matches and the number of invalid matches respectively, and the number of frames of Video1.

These plots can be summarized with Table 5.4, which contains the average number of extracted keypoints per frame (indicated with *key*), the average number of com-

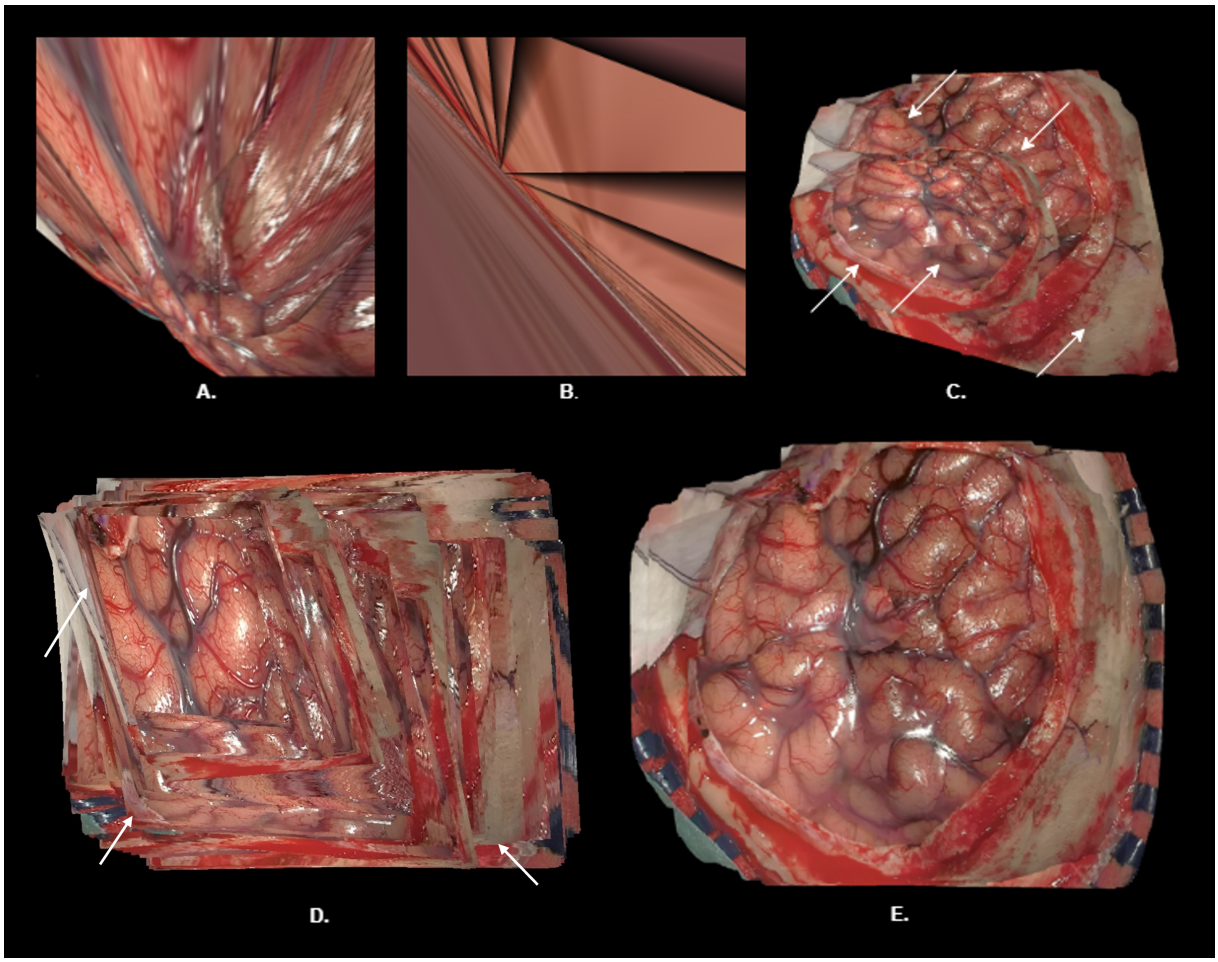


Figure 5.3: The figure shows the mosaics obtained with A.) BRISK, B.) ORB, C.) SIFT, D.) SuperPoint, E.) NeuroGlue. The white arrows indicates some inaccuracies and alignment errors in the registration. In this figure the results of video3 are reported.

puted matches per frame (*match*), the average number of invalid matches per frame (*inv_match*) and the percentage of invalid matches respect to the total (*inv_match*(%)) for the different tested methods (BRISK, ORB, SIFT, SuperPoint and NeuroGlue). The number of matches and invalid matches is referred to a single frame ($X_i$) but obviously it is related to a frame pair composed by $X_i$ and $X_{i-1}$.

Table 5.1: Mean ($m$) and variance ($\sigma$) of the structural similarity datasets ($SSIM$) computed for Video1 with BRISK, ORB, SIFT, SuperPoint and NeuroGlue methods. These values reflect the boxplot of Fig. 5.4.

| **5-frames*SSIM* Video 1** | BRISK | ORB | SIFT | SuperPoint | **NeuroGlue** |
|---|---|---|---|---|---|
| Mean ($m$) | 0.5093 | 0.4976 | 0.5105 | 0.3472 | **0.7557** |
| Variance ($\sigma$) | 0.0421 | 0.0416 | 0.0435 | 0.0365 | **0.0183** |

Table 5.2: Mean ($m$) and variance ($\sigma$) of the structural similarity datasets ($SSIM$) computed for Video2 with BRISK, ORB, SIFT, SuperPoint and NeuroGlue methods. These values reflect the boxplot of Fig. 5.4.

| **5-frames *SSIM* Video 2** | BRISK | ORB | SIFT | SuperPoint | **NeuroGlue** |
|---|---|---|---|---|---|
| Mean ($m$) | 0.5284 | 0.4957 | 0.7839 | 0.5478 | **0.7351** |
| Variance ($\sigma$) | 0.0387 | 0.0311 | 0.0111 | 0.0060 | **0.0136** |

Table 5.3: Mean ($m$) and variance ($\sigma$) of the structural similarity datasets ($SSIM$) computed for Video3 with BRISK, ORB, SIFT, SuperPoint and NeuroGlue methods. These values reflect the boxplot of Fig. 5.4.

| **5-frames *SSIM* Video 3** | BRISK | ORB | SIFT | SuperPoint | **NeuroGlue** |
|---|---|---|---|---|---|
| Mean ($m$) | 0.4858 | 0.4737 | 0.7361 | 0.5598 | **0.7611** |
| Variance ($\sigma$) | 0.0390 | 0.0567 | 0.0205 | 0.0161 | **0.0173** |

Table 5.4: The average number of extracted keypoints per frame ($key$), the average number of computed matches per frame pair ($match$), the average number of invalid matches per frame pair ($inv\_match$) and the percentage of invalid matches respect to the total ($inv\_match(\%)$) are reported for the investigated methods (BRISK, ORB, SIFT, SuperPoint and NeuroGlue) applied to Video1.

| | key | match | inv_match | inv_match(%) |
|---|---|---|---|---|
| BRISK | 1431.23 | 426.16 | 964.40 | 66.32 % |
| ORB | 390.85 | 237.22 | 153.59 | 39.30% |
| SIFT | 820.85 | 480.04 | 302.58 | 36.41% |
| SuperPoint | 75.97 | 67.16 | 8.81 | 11.64% |
| **NeuroGlue** | **596.06** | **549.96** | **45.30** | **7.64%** |

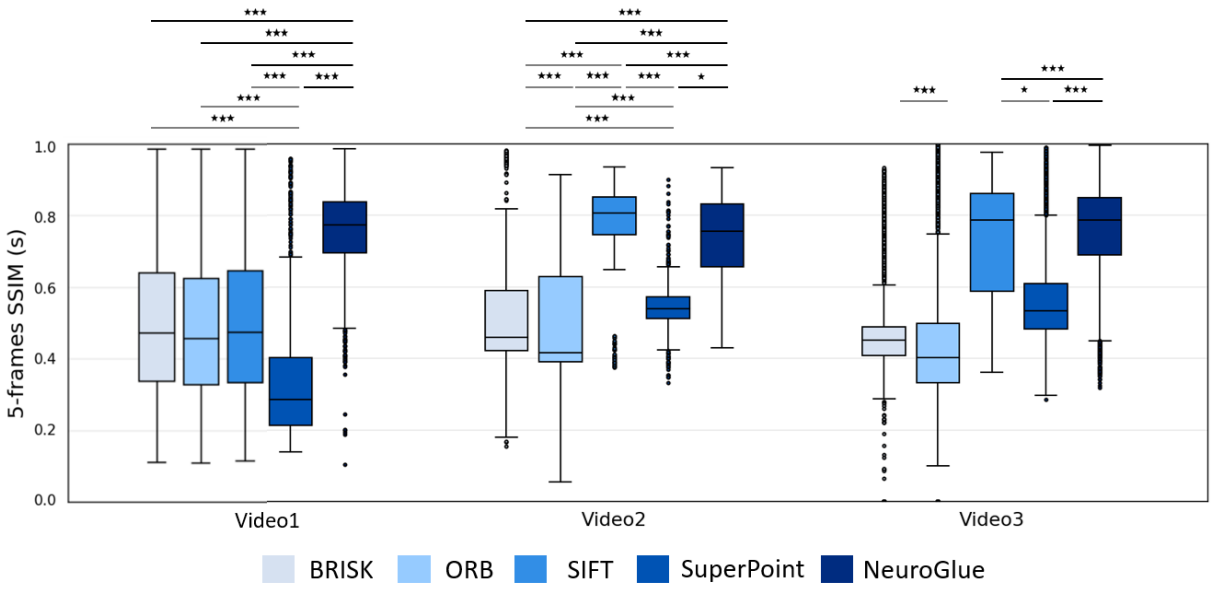


Figure 5.4: Boxplot of 5-frames *SSIM* for the tested methods: BRISK (in light grey), ORB (in light blue), SIFT (in blue), SuperPoint (in dark blue) and NeuroGlue (in night blue). The results of video1, video2 and video3 are reported. The stars indicate difference between the datasets from a statistical point of view. The number of stars is related to the obtained p-values and to the Wilcoxon Signed-Rank test results (Sec 4.3).

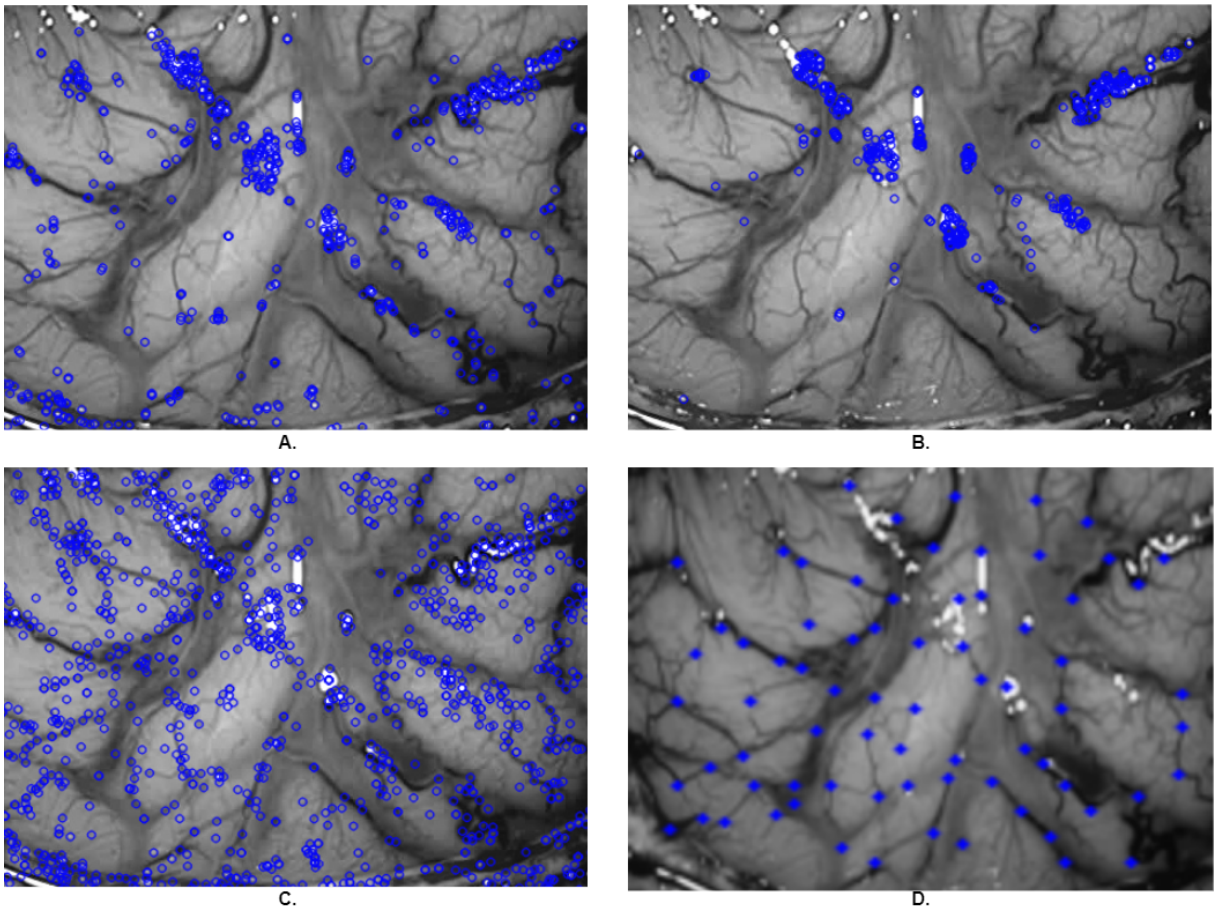


Figure 5.5: The figure shows the four keypoints representations obtained with A. BRISK, B. ORB, C.SIFT, D. NeuroGlue for frame0 of Video1.

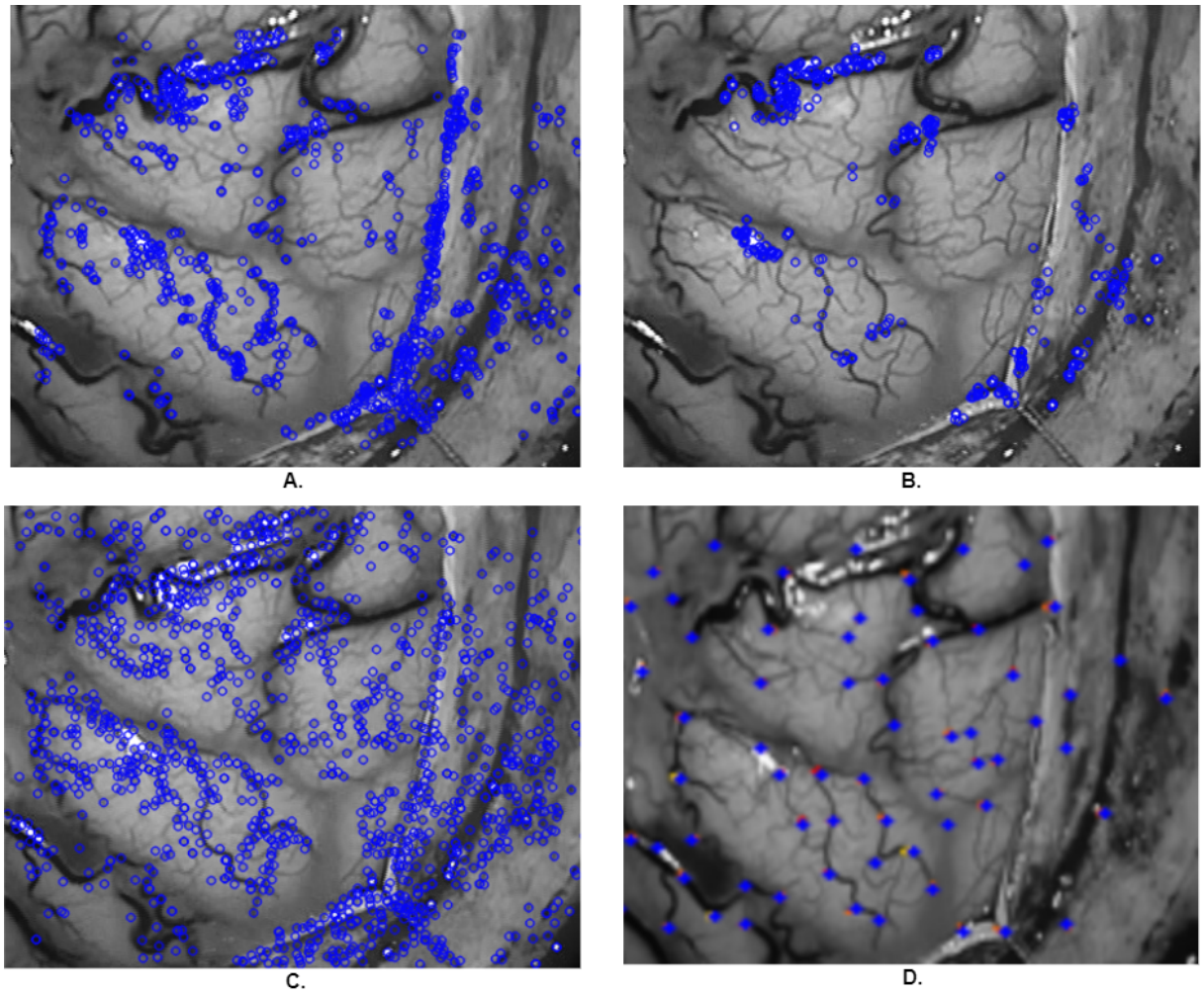


Figure 5.6: The figure shows the four keypoints representations obtained with A. BRISK, B. ORB, C.SIFT, D. NeuroGlue for frame884 of Video1.

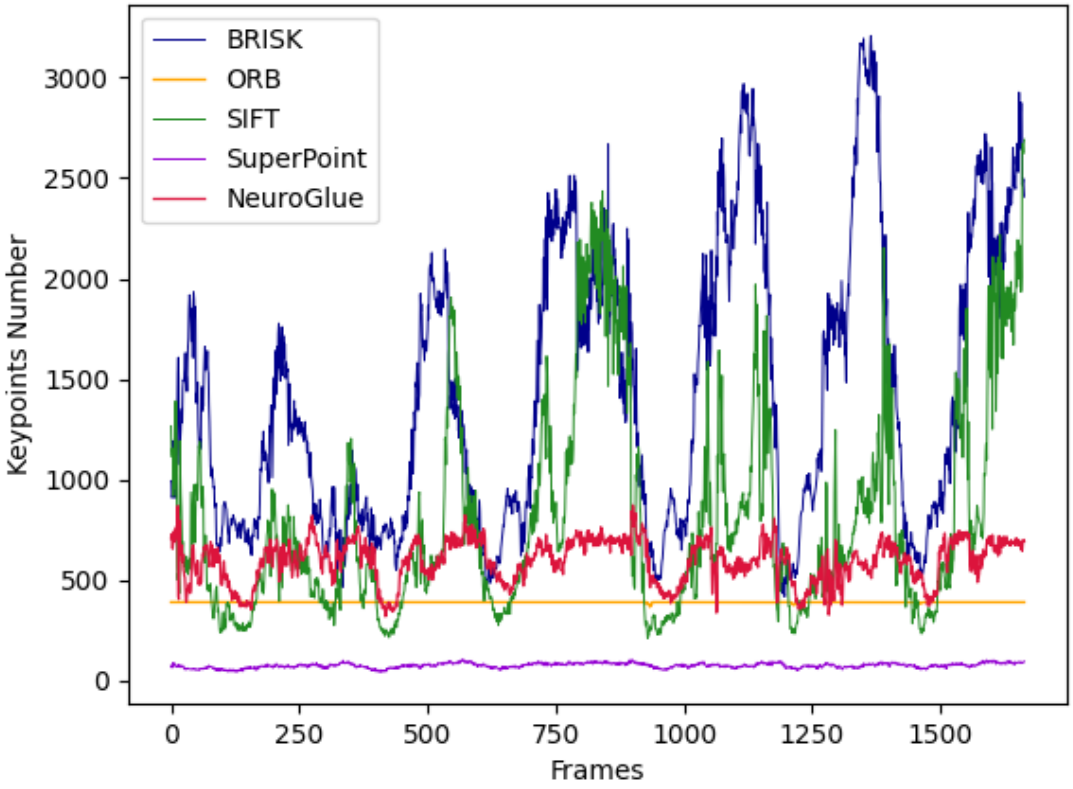


Figure 5.7: The plot illustrates the number of detected keypoints for each method: BRISK (in blue), ORB (in yellow), SIFT (in green), SuperPoint (in purple) and NeuroGlue (in red) respect to the number of frames. This example is reported for Video1.

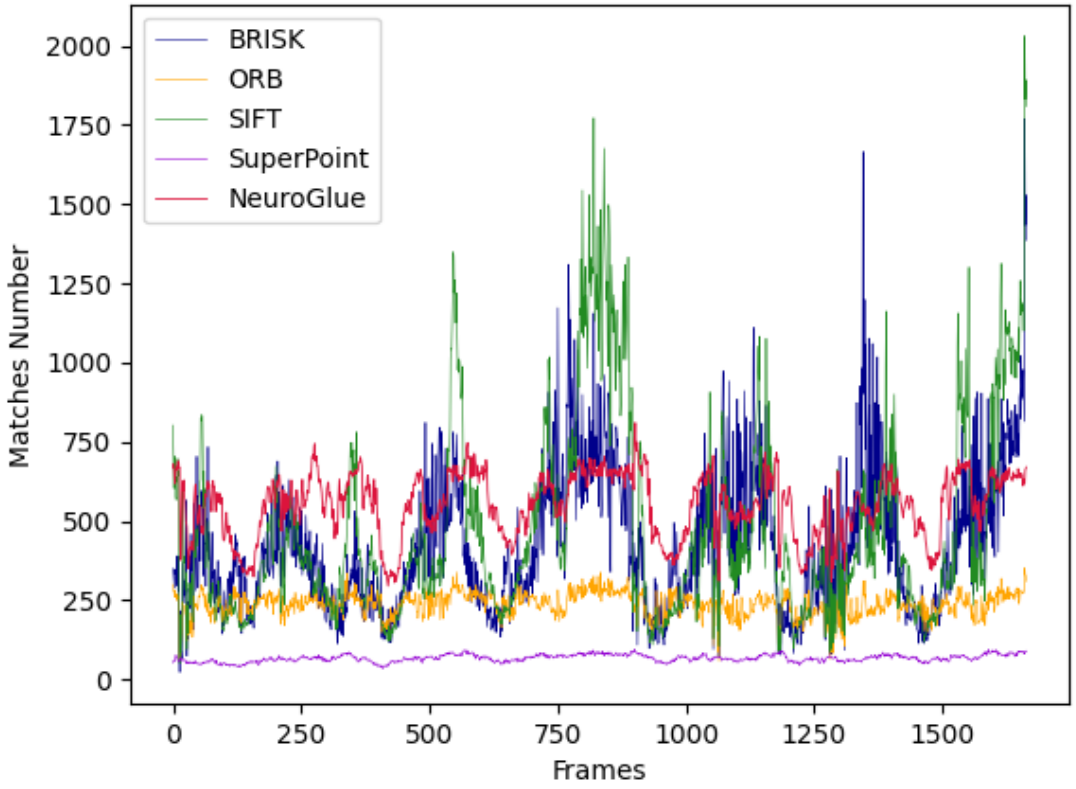


Figure 5.8: The plot illustrates the number of matches obtained with each method: BRISK (in blue), ORB (in yellow), SIFT (in green), SuperPoint (in purple) and NeuroGlue (in red) respect to the number of frames. This example is reported for Video1.

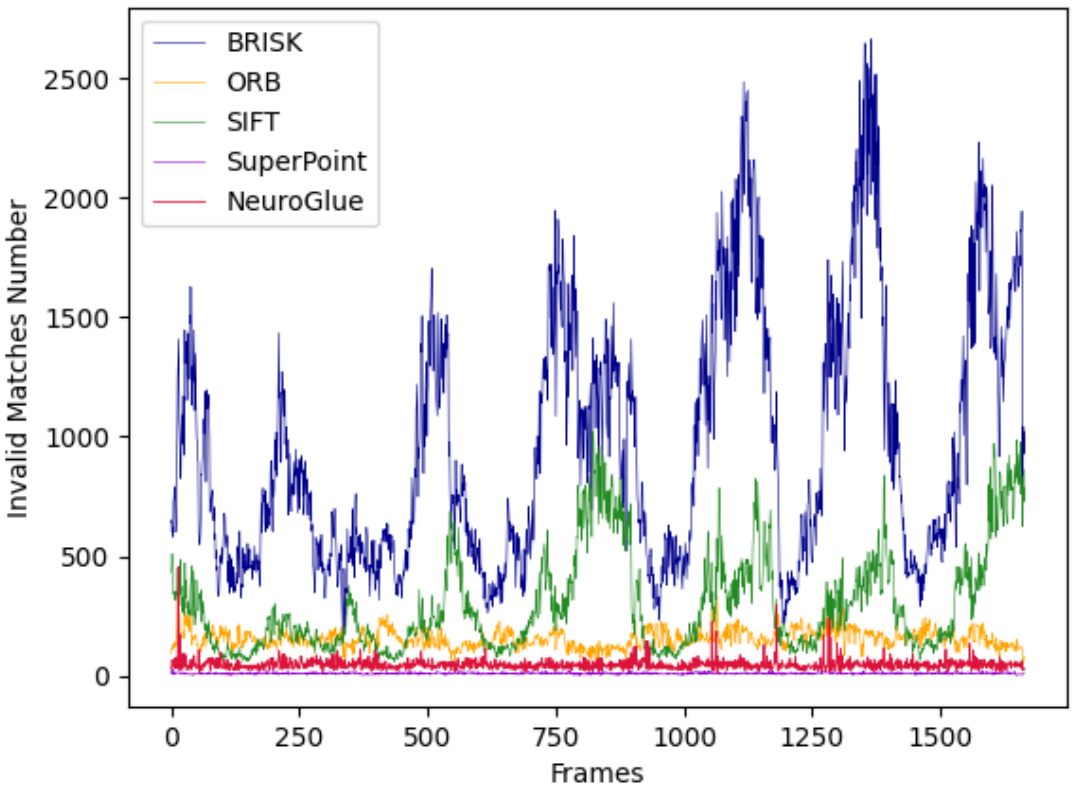


Figure 5.9: The plot illustrates the number of invalid matches obtained with each method: BRISK (in blue), ORB (in yellow), SIFT (in green), SuperPoint (in purple) and NeuroGlue (in red) respect to the number of frames. This example is reported for Video1.

# 6 | Discussion

A visual comparison of the mosaics, obtained with the different methods (BRISK, ORB, SIFT, SuperPoint and NeuroGlue), is reported in Fig. 5.1, for Video1, Fig. 5.2, for Video2 and Fig. 5.3, for Video3.

It is possible to observe that the classical algorithms lead to mosaics with several alignment errors and spatial distortions. This is particularly true for ORB keypoint detector: the reconstructions, shown in Fig. 5.1.B, Fig. 5.2.B and Fig. 5.3.B, are completely hidden by the huge image deformations.

Indeed the classical feature detectors may fail to extract enough keypoints into images which are characterized by repetitive patterns, illumination variation, and motion blur, typical elements of a neurosurgical dataset. These algorithms are primarily employed for the keypoints extraction of landscapes, buildings or everyday life objects.

In Fig. 5.5 and Fig. 5.6 two example of detected keypoints with BRISK ($A$), ORB ($B$), SIFT ($C$) and NeuroGlue ($D$) are present.

ORB is the feature detector which extracts the lowest number of keypoints (an average of 390 keypoints per frame as reported in Table 5.4) respect to the others traditional methods (BRISK and SIFT). This is confirmed also observing the keypoints number trend in Fig. 5.7. Accordingly ORB is the method which computes the less number of matches (on average 237) respect to BRISK and SIFT (Fig. 5.8).

This is a practical demonstration of one of the most relevant limitations of the classical features detectors: the reduced number of keypoints, detected with the traditional algorithm, restricts the space for matching. Indeed the same matching algorithm (KNN) is employed for all the classical features detectors and, for BRISK and SIFT, the number of matches is significantly higher (Fig. 5.8).

In general performing image registration with a limited number of keypoints and therefore matches, leads to inaccurate results.

On the other hand it is not important only the number of detected keypoints but also

their quality. In Fig. 5.5 and Fig. 5.6 it is evident that BRISK and SIFT extract a higher number of keypoints respect to NeuroGlue (and ORB) and it is shown also in Fig. 5.7. In particular BRISK detects an average of 1431 keypoints per frame and SIFT an average of 821 keypoints per frame respect to NeuroGlue (596 keypoints on average), as it is explained in Table 5.4.

However most of these keypoints is not employed in the matches generation. It is possible to confirm this statement observing Fig. 5.9, which illustrates the number of invalid matches trend during the registration process of Video1. BRISK and SIFT present the highest number of extracted keypoints but also the higher percentage of invalid matches, which is related to the number of rejected keypoints.

Indeed, with BRISK the 66.32% of the matches is rejected which corresponds to 964 matches and using SIFT the 39.30% of the matches is discarded (Table 5.4). Observing Fig. 5.5 and Fig. 5.6 the keypoints extracted with the classical algorithms, in particular with BRISK and SIFT, seem randomly positioned in the image. This characteristic underlines the generic behavior of this algorithms that are not used to deal with this typology of images. This explains why most of these keypoints are not employed for the matching phase.

In this way, although SIFT and BRISK produce a larger set of keypoints, the number of matches is on average lower respect to NeuroGlue: BRISK creates 426 matches on average per frame pair, SIFT 480 per frame pair and NeurGlue provides an average of 550 matches per frame pair, with a percentage of invalid matches of only 7.64% (Table 5.4).

The mosaic obtained with SuperPoint shows several errors in Fig. 5.1, for Video1, Fig. 5.2, for Video2 and Fig. 5.3, for Video3. Despite the promising keypoints detection network, the peculiar characteristics of neurosurgical images, as previously stated, and the lack of perfect coupling between the detection and matching phases provide inaccurate results. Indeed, the number of detected keypoints (76) and the number of computed matches (67) are very low respect to the other methods; it is evident in Table 5.4, in Fig. 5.7 and in Fig. 5.8. The SuperPoint architecture was previously trained with the COCO dataset [67], which includes images of everyday objects and landscapes. This behaviour underlines the network dependence on the training data.

The NeuroGlue low percentage of invalid matches (7.64%, Table 5.4) and the high number of computed matches (550, Table 5.4) demonstrates its ability to detect enough keypoints which are stable and adequate for image matching.

The Fully Convolutional Neural Network learned how to detect keypoints in images of a neurosurgical setting, characterized by regular patterns, illumination variations and motion blur. Indeed in Fig. 5.5 and Fig. 5.6 it is possible to observe that the keypoints positions follow the blood vessels present in the frame and are not randomly assigned.

The Attentional Graph Neural Network, characterized by the alternation of the Self Attentional Aggregation layer and the Cross Attentional Aggregation layer, gives a strong representation to the detected keypoints reducing the number of rejected ones for matching. The low percentage of invalid matches (on average equal to 7.64%, as reported in Table 5.4) is also related to the precise coupling between the keypoints detection FCNN and the Attentional Graph Neural Network for matching.

The quality of the obtained mosaics can be evaluated in terms of 5-frames Structural Similairty Metric (i.e. *SSIM*), as explained in Sec. 4.3. *SSIM* is computed during the registration process of Video1, Video2 and Video3 and the results are graphically represented in Fig. 5.4. The principal information (mean and variance of *SSIM*), that can be extracted from the boxplots of Fig. 5.4, are summarized in Table 5.1 for Video1 Table 5.2 for Video2 and Table 5.3 for Video3.

In general it is possible to assert that the quality of the results is reflected in the *SSIM* values. The mosaics obtained with ORB and BRISK are the most inaccurate and present huge frame distortions. These registration errors provide very low values of *SSIM*: the *SSIM* mean for ORB in Video1 is 0.4976 and the one of BRISK for the same video is 0.5093 (Table 5.1) and coherently they have the same behaviour also in Video2 (Table 5.2) and in Video3 (Table 5.3). Moreover they show the highest variance ($\sigma$) of *SSIM*: $\sigma$ related to ORB in Video1 is 0.0421 and related to BRISK of Video1 is 0.0416 and the same trend is present also in Video2 (Table 5.2) and in Video3 (Table 5.3).

A greater value of $\sigma$ is a demonstration of the low robustness of the considered method: the distribution of the *SSIM* values is not compact but variable. It means that the method is not able to maintain the stability achieved in the first moments of the registration procedure, during the whole process.

In general SIFT presents better results with respect to ORB and BRISK; indeed the mean value achieved in Video3 is 0.7361, a value comparable with the one obtained NeuroGlue (0.7611), as reported in Table 5.3. The reason is that SIFT is characterized by a better scale and rotation invariance, with respect to ORB and BRISK [22]. This is noticeable also observing the visual results in Fig. 5.1, for Video1, Fig. 5.2, for Video2 and Fig. 5.3, for Video3.

In particular in the first stages of the mosaicking development, SIFT results quite precise and shows its stability to scale and rotation. However going on with the registration (so using longer videos) the algorithm starts to fail and to be less accurate, as it is shown in Fig. 5.1.C for Video1 and Fig. 5.3.C for Video3, where the white arrows underline some errors.

In Video2 (Fig. 5.2.C) SIFT presents good results and it is shown also in *SSIM* values, which on average is 0.7361 as reported in Table 5.2. This behaviour is consistent with what was previously stated because Video2 is significantly shorter respect to Video1 and Video3: it has just 667 frames, while Video1 has 1677 frames and Video3 is composed by 3543 frames, as described in Sec. 3.6.

Considering a short video, so a reduced number of frames, the SIFT algorithm provides results quite precise due to its intrinsic characteristics. However, considering more frames and passing more times up to the same areas of the surgical field, underlines the problems of SIFT feature detector, which starts to lose the guide and introduce alignment errors in the reconstruction, as in Video1 and Video3.

This consideration justifies why the computed values of *SSIM* for Video2, present in Table 5.2, are higher for SIFT than for NeuroGlue. However observing in details Fig. 5.2.C (SIFT) and Fig. 5.2.E (NeuroGlue), the few alignment errors of NeuroGlue are located in the border, not in the brain tissue, as indicated by the white arrows in figure; instead SIFT shows reconstruction errors also in the middle of the surgical field. Errors at the border are not as relevant as those at the nervous tissue, because they don't belong to the interested field, involved in the surgical procedure.

The inaccurate coupling between SuperPoint and the KNN matching algorithm and the difficult challenge of dealing with neurosurgical data justify the reconstruction errors of the panoramas and this behaviour is reflected in *SSIM* calculation (Fig. 5.4). Indeed the *SSIM* mean, achieved with SuperPoint is 0.3472 (Table 5.1) for Video 1, 0.5478 (Table 5.2) for Video2 and 0.5598 for Video3 (Table 5.3).

NeuroGlue panorama stands out clearly from the other methods, this is due to the stability and the strength of the network which learned to deal with neurosurgical images, thanks to the adapted training (described in Sec. 4.1).

The higher accuracy is shown in the higher mean of *SSIM*, and the robustness is demonstrated with the lower variance of *SSIM*. For example for Video3, the *SSIM* mean is 0.7611 and $\sigma$ is 0.0173 (Table 5.3). These values and the ones obtained with Video1 (Table 5.1) and Video2 (Table 5.2), demonstrate that NeuroGlue outperforms the clas-

sical algorithms (BRISK ORB and SIFT) and also the learning based keypoints detection realized with SuperPoint coupled with KNN.

The Wilcoxon Signed-Rank test is performed, as Fig. 5.4 indicates, and it underlines a significant difference between NeuroGlue and the other methods.

# 7 | Conclusion

This thesis presents a learning based mosaicking framework, able to build a real-time panorama of a neurosurgical environment. To the best of our knowledge, the presented algorithm is the first attempt to introduce the mosaicking tool in neurosurgery to overcome the low visibility issue for the SM magnifications performed during surgery.

The peculiar conditions of a surgical setting could affect the quality of the mosaicking result. These issue introduces the need to find stable keypoints detectors and establish stronger connections between the keypoints of consecutive frames. These aspects guarantee a greater stability and robustness, essential characteristics for the correct homography matrix estimation and the consequent mosaicking accurate development.

The proposed method showed to achieve better performance in terms of *SSIM* compared to the traditional feature detection algorithms (BRISK, ORB and SIFT) and also respect to the SuperPoint method, underlying the importance of the domain adaptation procedure, described in Sec. 4.1 and of the Attentional Graph Neural Network stability (Sec. 3.2).

The real-time mosaic development and thus the introduction of an expanded view of the surgical theatre could represent a valuable tool to deal with low visibility in neurosurgery and to tackle with challenging tumors or lesions localization for the navigation systems inaccuracies.

This expansion tool can be comfortably used during the neurosurgical procedures without the need to introduce a new sensor or a new device respect to the ones already present in the operating room, such as the SM and the neuro-navigation systems.

Indeed the surgeon can work on the brain anatomy, observing each details, thanks to the high magnifications of the SM, and at the same time he is able to have a broader view of the entire scene, without further moving the microscope, storing the hand-held instruments or changing the magnifications during the procedure, reducing significantly the timing of surgery.

In this way the obtained panorama can represent an important reference for the surgeon

in the operating room.

This tool could be involved in the surgical procedures thanks to its easy application for surgeons and because its employment is almost independent from surgeons actions or from the occurrence of unexpected events such as a sudden movement of the SM.

The proposed mosaicking framework is able to recognize the presence of an incorrect and unwanted movement of the SM. In particular the current architecture involves the homography estimation, decomposition and analysis for each extracted frame. From this homography investigation, the invalid frames related to the unwanted movements are identified and discarded. In this way the panorama recorded before the unexpected event is restored without the need to start a new registration.

A limitation of this work is that NeuroGlue was trained and tested on a reduced dataset: only three videos were available. In order to reinforce its validation it is needed to expand the neurosurgical dataset, providing more videos of the neurosurgical setting, including also videos with a sudden rapid movement of the miscroscope to verify the filtering operation.

Moreover the NeuroGlue based mosaicking was applied only for the reconstruction of the brain superficial layers. However the obtained results are promising and give hope that the proposed method can also be employed for the reconstruction of deeper anatomical structures.

This project's future development and application could include its effective integration with the intra-operative neuro-navigation systems output. This integration could minimize the inaccuracies for lesions or tumour localization caused by the brain shift phenomenon.

The obtained panoramas could be integrated also with the pre-operative images, such as the MRI, in order to create an anatomical and functional biomedical image which includes more information respect to the single MRI.

# Bibliography

[1] Silvia Lanfranconi, Elisa Scola, Giulio Andrea Bertani, Barbara Zarino, Roberto Pallini, Giorgio d'Alessandris, Emanuela Mazzon, Silvia Marino, Maria Rita Carriero, Emma Scelzo, et al. Propranolol for familial cerebral cavernous malformation (treat_ccm): study protocol for a randomized controlled pilot trial. *Trials*, 21(1):1–10, 2020.

[2] Matteo Malinverno, Claudio Maderna, Abdallah Abu Taha, Monica Corada, Fabrizio Orsenigo, Mariaelena Valentino, Federica Pisati, Carmela Fusco, Paolo Graziano, Monica Giannotta, et al. Endothelial cell clonal expansion in the development of cerebral cavernous malformations. *Nature communications*, 10(1):1–16, 2019.

[3] Daniel A Snellings, Courtney C Hong, Aileen A Ren, Miguel A Lopez-Ramirez, Romuald Girard, Abhinav Srinath, Douglas A Marchuk, Mark H Ginsberg, Issam A Awad, and Mark L Kahn. Cerebral cavernous malformation: from mechanism to therapy. *Circulation research*, 129(1):195–215, 2021.

[4] Issam A Awad and Sean P Polster. Cavernous angiomas: deconstructing a neurosurgical disease: Jnspg 75th anniversary invited review article. *Journal of neurosurgery*, 131(1):1–13, 2019.

[5] Kelly D Flemming and Giuseppe Lanzino. Cerebral cavernous malformation: what a practicing clinician should know. In *Mayo Clinic Proceedings*, volume 95, pages 2005–2020. Elsevier, 2020.

[6] Pretty Sara Idiculla, Dhineshreddy Gurala, Jobin Philipose, Kartikeya Rajdev, and Prateek Patibandla. Cerebral cavernous malformations, developmental venous anomaly, and its coexistence: a review. *European Neurology*, 83(4):360–368, 2020.

[7] Shengchao Xu, Lu Tang, Xizhe Li, Fan Fan, and Zhixiong Liu. Immunotherapy for glioma: current management and future application. *Cancer letters*, 476:1–12, 2020.

[8] Mary Elizabeth Davis. Glioblastoma: overview of disease and treatment. *Clinical journal of oncology nursing*, 20(5):S2, 2016.

[9] Gaurav Shukla, Gregory S Alexander, Spyridon Bakas, Rahul Nikam, Kiran Talekar, Joshua D Palmer, and Wenyin Shi. Advanced magnetic resonance imaging in glioblastoma: a review. *Chin Clin Oncol*, 6(4):40, 2017.

[10] Ricky Chen, Matthew Smith-Cohn, Adam L Cohen, and Howard Colman. Glioma subclassifications and their clinical significance. *Neurotherapeutics*, 14(2):284–297, 2017.

[11] Richard M Young, Aria Jamshidi, Gregory Davis, and Jonathan H Sherman. Current trends in the surgical management and treatment of adult glioblastoma. *Annals of translational medicine*, 3(9), 2015.

[12] Qing Lan, Michael Sughrue, Nikolai J Hopf, Kentaro Mori, Jaechan Park, Hugo Andrade-Barazarte, Mangaleswaran Balamurugan, Macro Cenzato, Giovanni Broggi, Dezhi Kang, et al. International expert consensus statement about methods and indications for keyhole microneurosurgery from international society on minimally invasive neurosurgery. *Neurosurgical review*, 44(1):1–17, 2021.

[13] Ling Ma and Baowei Fei. Comprehensive review of surgical microscopes: technology development and medical applications. *Journal of biomedical optics*, 26(1):010901, 2021.

[14] Judith Rösler, Stefan Georgiev, Anna L Roethe, Denny Chakkalakal, Güliz Acker, Nora F Dengler, Vincent Prinz, Nils Hecht, Katharina Faust, Ulf Schneider, et al. Clinical implementation of a 3d4k-exoscope (orbeye) in microneurosurgery. *Neurosurgical Review*, 45(1):627–635, 2022.

[15] Inês Machado, Matthew Toews, Jie Luo, Prashin Unadkat, Walid Essayed, Elizabeth George, Pedro Teodoro, Herculano Carvalho, Jorge Martins, Polina Golland, et al. Non-rigid registration of 3d ultrasound for neurosurgery using automatic feature detection and matching. *International journal of computer assisted radiology and surgery*, 13(10):1525–1538, 2018.

[16] Tsukasa Koike, Taichi Kin, Shota Tanaka, Yasuhiro Takeda, Hiroki Uchikawa, Taketo Shiode, Toki Saito, Hirokazu Takami, Shunsaku Takayanagi, Akitake Mukasa, et al. Development of innovative neurosurgical operation support method using mixed-reality computer graphics. *World neurosurgery: X*, 11:100102, 2021.

[17] Loic Peter, Marcel Tella-Amo, Dzhoshkun Ismail Shakir, Jan Deprest, Sebastien Ourselin, Juan Eugenio Iglesias, and Tom Vercauteren. Active annotation of informative overlapping frames in video mosaicking applications. *arXiv preprint arXiv:2012.15343*, 2020.

[18] Shaharyar Ahmed Khan Tareen and Zahra Saleem. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In *2018 International conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–10. IEEE, 2018.

[19] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[20] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021.

[21] John Vourvoulakis, John Kalomiros, and John Lygouras. Fpga-based architecture of a real-time sift matcher and ransac algorithm for robotic vision applications. *Multimedia Tools and Applications*, 77(8):9393–9415, 2018.

[22] Dibyendu Mukherjee, QM Jonathan Wu, and Guanghui Wang. A comparative experimental study of image feature detectors and descriptors. *Machine Vision and Applications*, 26(4):443–466, 2015.

[23] Hans P Moravec. Techniques towards automatic visual obstacle avoidance. 1977.

[24] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[25] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, International Journal of Computer Vision, 1991.

[26] Friedrich Heitger, Lukas Rosenthaler, Rüdiger Von Der Heydt, Esther Peterhans, and Olaf Kübler. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision research*, 32(5):963–981, 1992.

[27] Wolfgang Förstner. A framework for low level feature extraction. In *European Conference on Computer Vision*, pages 383–394. Springer, 1994.

[28] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994.

[29] Stephen M Smith and J Michael Brady. Susan—a new approach to low level image processing. *International journal of computer vision*, 23(1):45–78, 1997.

[30] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Comparing and evaluating

interest points. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 230–235. IEEE, 1998.

[31] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of computer vision*, 37(2):151–172, 2000.

[32] Daniela Hall, Bastian Leibe, and Bernt Schiele. Saliency of interest points under scale changes. In *BMVC*, volume 2, pages 646–655, 2002.

[33] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[34] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 525–531. IEEE, 2001.

[35] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[36] Manuel G Forero, Claudia L Mambuscay, María F Monroy, Sergio L Miranda, Dehyro Méndez, Milton Orlando Valencia, and Michael Gomez Selvaraj. Comparative analysis of detectors and feature descriptors for multispectral image matching in rice crops. *Plants*, 10(9):1791, 2021.

[37] G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2, 2004.

[38] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

[39] Yue Feng, Jinchang Ren, Jianmin Jiang, Martin Halvey, and Joemon M Jose. Effective venue image retrieval using robust feature extraction and model constrained matching for mobile robot localization. *Machine Vision and Applications*, 23(5):1011–1027, 2012.

[40] Jian Gao, Xinhan Huang, and Bo Liu. A quick scale-invariant interest point detecting approach. *Machine Vision and Applications*, 21(3):351–364, 2010.

[41] Kaiyang Liao, Guizhong Liu, and Youshi Hui. An improvement to the sift descriptor for image representation and matching. *Pattern Recognition Letters*, 34(11):1211–1220, 2013.

[42] Pierre Moreels and Pietro Perona. Evaluation of features detectors and descriptors based on 3d objects. *International journal of computer vision*, 73(3):263–284, 2007.

[43] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[44] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[45] David Nistér and Henrik Stewénius. Linear time maximally stable extremal regions. In *European conference on computer vision*, pages 183–196. Springer, 2008.

[46] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[47] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European conference on computer vision*, pages 430–443. Springer, 2006.

[48] Features from accelerated segment test, Dec 2021.

[49] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. Censure: Center surround extremas for realtime feature detection and matching. In *European conference on computer vision*, pages 102–115. Springer, 2008.

[50] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010.

[51] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[52] Franco Hidalgo and Thomas Bräunl. Evaluation of several feature detectors/extractors on underwater images towards vslam. *Sensors*, 20(15):4343, 2020.

[53] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, pages 2548–2555. Ieee, 2011.

[54] Cuiyin Liu, Jishang Xu, and Feng Wang. A review of keypoints' detection and feature description in image registration. *Scientific Programming*, 2021, 2021.

[55] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *2012 IEEE conference on computer vision and pattern recognition*, pages 510–517. Ieee, 2012.

[56] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *European conference on computer vision*, pages 214–227. Springer, 2012.

[57] Pablo F Alcantarilla and T Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7):1281–1298, 2011.

[58] Kyi Pyar Win and Yuttana Kitjaidure. Biomedical images stitching using orb feature based approach. In *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, volume 3, pages 221–225. IEEE, 2018.

[59] Szymon Grabowski and Tomasz M Kowalski. Algorithms for all-pairs hamming distance based similarity. *Software: Practice and Experience*, 51(7):1580–1590, 2021.

[60] Edgar Fabian Aguilar Calzadillas. *Sparse Stereo Visual Odometry with Local Non-Linear Least-Squares Optimization for Navigation of Autonomous Vehicles*. PhD thesis, Carleton University, 2019.

[61] Siddique Abu Bakar, Xiaoming Jiang, Xiangfu Gui, Guoquan Li, and Zhangyong Li. Image stitching for chest digital radiography using the sift and surf feature extraction by ransac algorithm. In *Journal of Physics: Conference Series*, volume 1624, page 042023. IOP Publishing, 2020.

[62] Frazer K Noble. Comparison of opencv's feature detectors and feature matchers. In *2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pages 1–6. IEEE, 2016.

[63] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[64] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):105–119, 2008.

[65] Vincent Lepetit. Tilde: A temporally invariant learned detector. In *Conference on Computer Vision and Pattern Recognition*, pages 1–10. ., 2015.

[66] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016.

[67] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[68] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018.

[69] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018.

[70] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Self-improving visual odometry. *arXiv preprint arXiv:1812.03245*, 2018.

[71] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

[72] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

[73] Nidhal K EL Abbadi, Safaa Alwan Al Hassani, and Ali Hussein Abdulkhaleq. A review over panoramic image stitching techniques. In *Journal of Physics: Conference Series*, volume 1999, page 012115. IOP Publishing, 2021.

[74] Sophia Bano, Francisco Vasconcelos, Marcel Tella-Amo, George Dwyer, Caspar Gruijthuijsen, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Jan Deprest, and Danail Stoyanov. Deep learning-based fetoscopic mosaicking for field-of-view expansion. *International journal of computer assisted radiology and surgery*, 15(11):1807–1816, 2020.

[75]

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **2D** | Two-Dimensional |
| **3D** | Three-Dimensional |
| ***Acc*** | Accuracy |
| **AGNN** | Attentional Graph Neural Network |
| **AI** | Artificial Intelligence |
| **AKAZE** | Accelerated-KAZE |
| **ANN** | Artificial Neural Network |
| **BF** | Brute Force |
| **BN** | Batch Normalization |
| **BRIEF** | Binary Robust Independent Elementary Features |
| **BRISK** | Binary Robust Invariant Scalable Key-points |
| **CAI** | Computer Assisted Intervention |
| **CCM** | Cerebral Cavernous Malformation |
| **CENSURE** | Center Surround Extremas |
| **CNN** | Convolutional Neural Network |
| **CT** | Computed Tomography |
| **DL** | Deep Learning |
| **DoG** | Difference-of-Gaussians |
| **FAST** | Features from Accelerated Segment Test |
| **FCNN** | Fully Convolutional Neural Network |
| **FLANN** | Fast Library for Approximate Nearest Neighbors |
| **FoV** | Field of View |
| **FREAK** | Fast Retina Key-point |
| **GAN** | Generative Adversarial Network |
| **KNN** | K-Nearest Neighbors |

| | |
|---|---|
| **LIFT** | Learn Invariant Feature Transform |
| **LM** | Levenberg-Marquardt |
| **MIS** | Minimally Invasive Surgery |
| **MLP** | Multi-Layer Perceptron |
| **MRI** | Magnetic Resonance Imaging |
| **MSER** | Maximally Stable Extremal Regions |
| **ORB** | Oriented FAST and Rotated BRIEF |
| **PCA** | Principal Component Analysis |
| **PCA-SIFT** | Principal Component Analysis-SIFT |
| ***Prec*** | Precision |
| **RANSAC** | RANdom Sample Consensus |
| **RoI** | Region of Interest |
| **SfM** | Structure-from-Motion |
| **SGD** | Stochastic Gradient Descent |
| **SLAM** | Simultaneous Localization and Mapping |
| **SIFT** | Scale Invariant Feature Transform |
| **SM** | Surgical Microscope |
| **SSIM** | Structural Similarity Measure |
| **SURF** | Speeded Up Robust Features |
| **SVD** | Singular Value Decomposition |
| **SVM** | Support Vector Machine |
| **TILDE** | Temporally Invariant Learned Detector |
| **TTTS** | Twin-to-Twin Transfusion Syndrome |
| **VGG** | Visual Geometry Group |
| **WHO** | World Health Organization |