CentraleSupélec

3D localisation from multi-view X-ray projections

Pôle projet Intelligence Artificielle

Réalisé par : *P10.S7.06*
Massine El Khader
Riccardo Zanchetta
Thomas Duflos
João Pedro Regazzi
Khalil Ben Gamra

13 mars 2024

Table des matières

1 Introduction

1.1 Our client

This project was proposed by SAFRAN, a tier-1 aerospace company that is developing technologies that are meant to enable carbon-free air transportation. Since this corporate R&T entity produces complex aeronautic pieces, assisted defect inspection from multi-view x-ray imaging can reduce cost while maintaining high quality standards of SAFRAN.

The international high-technology group SAFRAN works with propulsion, equipment, and interiors, along with operations in space and defense sectors. With the goal to contribute to a safer world, SAFRAN tries to make air transport more environmentally friendly, comfortable, and accessible. The company has a workforce of 83 thousand employees, has presence in all continents and achieved a turnover of 19 billion euros in 2022. Either independently or through strategic partnerships, it holds leading global and European positions in its markets. The commitment to research and development initiatives that align with environmental priorities is at its core.

Additionally, SAFRAN's governance structure operates under French law, is listed on NYSE Euronext Paris and underscores its significance in the business landscape. In this context, the company is included in the CAC 40 and in the Euro Stoxx 50 indexes. The Executive Committee, representative of SAFRAN's diverse activities, executes operational strategies aligned with the directives set by the Board of Directors. This governance model ensures a balanced distribution of powers, allowing SAFRAN to maintain agility in response to the dynamic economic, financial, and competitive landscape in which it operates. Through the dedication of its workforce, a commitment to innovation, and a focus on operational excellence, SAFRAN continues to pioneer and support high-tech solutions, contributing significantly to a safer world and enhancing the sustainability, comfort, and accessibility of air transport, while extending its expertise to critical areas such as defense and space exploration.

1.2 The problem

In the domain of artificial intelligence and computer vision, the pursuit of accurate image recognition is a well-known problem. Traditional models that embed indicators into images usually only use 2D data and do not consider multiple points of view of objects. This report tackles the challenge of accurate 3D localization from multi-view projections and goes into a comprehensive exploration of the state-of-the-art methodologies, analyzing recent advancements in the field. In this context, the problem to solve is to find and implement the best architecture to address the challenge of developing a model to set indicators into 3D objects, using its geometry and 10 different point-of-view images as input.

Our journey began with an in-depth study of state-of-the-art architectures for 3D detection. Team members researched papers covering various approaches, as presented in section two. This collaborative effort laid the groundwork for subsequent phases. Subsequently, we focused on dataset exploration, leading us to the LIDC and nuScene

datasets. The first is a cornerstone in lung cancer detection research, and the second is rich in labeled images and additional sensor sweeps. Both provided diverse and complex data for training and testing.

Our exploration led to the development of multiple approaches. The MV2D approach for multi-view 3D detection posed coding challenges that, despite our efforts, are not yet solved. Transitioning to 2D detection, we pursued a Faster R-CNN tuning approach, which involved fine-tuning the model with the LIDC dataset, demanding a deep understanding of lung image analysis. Finaly, our third implementation approach was the Geometry-based 3D detection, emphasizing the fusion of deep learning and geometric insights. Both Faster R-CNN and Geometry-based 3D detection approaches ran successfully and had their results measured. In this regard, our measure of success centered around Intersection over Union (IoU), a metric capturing the overlap between predicted and ground truth bounding boxes. This provided a quantitative assessment of model accuracy in localization tasks.

In summary, the main goal of this report is to pave the way for the next students that will continue this project, presenting all its essential information. In this context, we display our planning with the clients, our path analyzing six state-of-the-art approaches, our databases, our reasoning for the choice of the implemented models, their training and testing processes, their results and, finaly, our conclusions.

2 State-of-the-art

2.1 A Light Touch Approach to Teaching Transformers Multi-view Geometry

The central idea presented in the discussed paper revolves around enhancing the capability of Transformer models in understanding multi-view geometry for visual tasks. Transformers, as we've seen in the previous section, is known for their proficiency in learning from visual data and often lack manually-specified priors, making them highly flexible. However, this flexibility becomes a challenge when dealing with multi-view geometry.

The complexity of multi-view geometry comes from the infinite possible variations in 3D shapes and viewpoints, requiring a system that is both adaptable (due to the variety of possible views) and precise (as projective geometry follows rigid laws). The paper proposes a novel approach, termed a "light touch" method, to address this. The idea is to guide Transformers to learn multi-view geometry, allowing them to deviate when necessary.

The guiding mechanism involves the use of epipolar lines but only during the training phase of the Transformer model. In the context of multi-view geometry, epipolar lines represent the potential locations in one image where a point from another image could be found, given the relative positions of the cameras. By focusing the Transformer's attention on these lines, the model is encouraged to consider geometrically plausible areas for matching points across different views.

The unique aspect of this approach is that it does not rigidly bind the Transformer to epipolar constraints but rather uses them as a soft guide. The attention mechanism of the Transformer is adjusted so that it pays more attention to regions along the epipolar lines during the learning process. This is achieved by penalizing attention to areas outside these lines and rewarding attention within them.

An advantage of this method is its independence from camera pose information during actual usage, unlike some previous techniques. This makes the method more versatile and practical for real-world applications. The paper demonstrates the effectiveness of this approach, especially in scenarios like pose-invariant object recognition, where traditional Transformer networks struggle due to significant viewpoint differences between query and retrieved images. The disadvantage of this method is that it's not really adapted to detecting the position of an object.

In summary, the paper presents an innovative approach to use Transformers with an understanding of multi-view geometry, enhancing their performance in visual tasks involving complex spatial relations. This is achieved by a training method that leverages epipolar geometry as a guiding principle, while still maintaining the inherent flexibility of the Transformer model.

2.2 Quad-DIP for X-ray cargo image decomposition

The article introduces a novel unsupervised image decomposition method called Quad-DIP, designed to accurately separate vehicle structures and cargo information

in X-ray scanned images of cargo vehicles. The main challenge arises from the lack of ground truth data for empty vehicles and the complex nature of vehicle components. To address this, the proposed method employs a dual-stage decomposition network, through a double-DIP (DDIP) structure which has two *deep_image_prior* (DIP) blocks, one for vehicle image extraction and other for cargo image extraction.

Additionally, each stage of the Quad-DIP has cross-mixing blocks so that the pairs of extracted images of vehicle structure and cargo are changed. The first phase adds different goods and vehicles structures together while the second one does the mixing again, reconstructing the initial image. This innovative strategy proportionates a performance enhancement on the decomposition of the vehicle structure in X-ray images.

Furthermore, the loss function incorporates reconstruction and similarity losses to ensure precise decomposition and consistency across stages. Since the original image is reconstructed at the end of the process, the difference between the reconstruction and the input image is considered as measurement of error. Moreover, the extracted vehicle structure images should be the same and their difference is also seen as a measurement for the loss function. In this context, it's possible to testify that this approach to image decomposition has the constraints of only working with vehicles of the same structure and with images from the same point-of-view.

Ultimately, Quad-DIP method outperforms existing methods in accurately decomposing X-ray cargo vehicle images, making it a promising solution, especially in scenarios where empty vehicle images are unavailable for supervision. On the other hand, it's not clear how we could implement it for our project, given its constraints.

2.3 Multi-view Vision Transformers for Object Detection

Mutli-view object detection refers to localize a specific object with multiple image of the same scene [2]. There are no other specific constraints on the images : image points can overlap, or the object may not be clearly visible in one of the images. To talk about this architecture, it's important to explain how modern detectors are built : with three subnetworks (backbone, neck and head). The purpose of the backbone is to extract the feature map. Some detectors use a sub-network –the neck– used to group together the features of the different layers of the backbone. Finally, the purpose of the head detector is to locate objects according to the features extracted from the backbone. In this work, the authors use the combination of intermediates features from multiple views (get from the backbone of standard detection architecture, as ViT). These intermediates features are combine with a transformers based architecture. A Transformers decoder use the target views feature vectors and the feature from a source view to fusion all those features. Thus, the feature representation is linked to the information from other views, making it consider the 3D scene geometry. We apply a such Transformer decoder to each view, thereby all views are target and source. If there is more than two views, we simply consider the feature map of each view with concatenation. The authors have considered three different object detection architecture(YOLOX, Deformable DETR, Swin Transformers) where there model is integrated. In all cases MVViT could improve object detection compared to single view object detection. For a multi-

camera surveillance dataset there is an improvement of 0.7% of the performance for both COCO AP and COCO $AP_{0.5}$, and for an X-ray security imagery dataset, the results are even greater : +3.0% COCO AP and +1.9% COCO $AP_{0.5}$.

2.4 Enhanced Defect Localization in X-Ray Images through Integrated Geometric Information

This paper delves into the intricacies of a computer vision and machine learning project aimed at developing a robust pipeline for the accurate detection and localization of defects in three-dimensional (3D) objects using X-ray images. The proposed methodology not only employs state-of-the-art deep learning architectures but also integrates geometric information, offering a comprehensive approach to enhance defect identification. This document provides an in-depth exploration of the project structure, outlining the key techniques and methodologies employed.

The project's primary goal is to leverage computer vision techniques for defect detection in complex structures using X-ray images. Our approach involves a carefully designed pipeline that seamlessly integrates geometric information into the analysis, addressing limitations observed in conventional methods. The project's significance lies in its potential to revolutionize defect identification processes, especially in industries where precision is paramount.

2.5 Multi-view X-Ray R-CNN

The authors propose a CNN-based object detection approach for multi-view X-ray image data. They introduce a novel multi-view pooling layer to perform a 3D aggregation of 2D CNN-features extracted from each view, ensuring geometric consistency. They also introduce an end-to-end trainable multi-view detection pipeline based on Faster R-CNN. The approach shows significant accuracy gains compared to single-view detection while being more efficient. The authors extend the Faster R-CNN method to multi-view X-ray images using late fusion and introduce a novel multi-view pooling layer for feature aggregation.

The MX-RCNN approach combines feature extraction on 2D X-ray images and multi-view pooling to detect objects in a 3D feature space. A ResNet-50 architecture is used for feature extraction, and a multi-view pooling layer combines feature maps into a common 3D feature volume. The RPN proposes 3D regions which are then classified and regressed to determine 3D bounding box parameters. K-means clustering is used to optimize the anchor boxes, and the multi-view pooling layer maps feature maps from different views into a 3D feature volume using known X-ray image formation geometry. Different variants of the multi-view pooling layer are implemented : MX-RCNN avg and MX-RCNN max. The IoU thresholds are converted to account for differences in 3D and 2D spaces.

An example plot is shown to illustrate the relevant beams in the multi-view pooling of a specific output cell. The geometry of the bounding box is slightly different. The text also discusses the thresholds for 2D and 3D bounding boxes and the computational

cost of the method. It mentions the dataset that was used, which consists of X-ray recordings of hand luggage. Different subsets of the dataset are described, along with the annotations available. The process of generating 3D bounding box annotations from 2D bounding boxes is explained.

MX-RCNN is a multi-view object detection pipeline for X-ray images that combines features from multiple views to create a 3D representation of the object. The system outperforms single-view detection methods and is computationally efficient. The approach is robust to missing views and relies on all views to construct the 3D feature representation.

The text discusses various references related to object detection and recognition using X-ray imagery. The references cover topics such as transfer learning, multi-view object detection, deep learning techniques, and 3D object detection. These references provide insights into different approaches and methods for improving the accuracy and efficiency of object detection in X-ray images.

2.6 Object as Query : Lifting any 2D Object Detector to 3D Detection

The paper introduces the Multi-View 2D Objects guided 3D Object Detector (MV2D) as a framework designed to enhance the localization of objects in multi-view images. The motivation for MV2D stems from the advancements in 2D object detection methods and the desire to leverage these developments for 3D object detection. MV2D's pipeline involves extracting image feature maps from multi-view images and applying a 2D object detector to obtain 2D object bounding boxes. These bounding boxes are used to generate dynamic object queries, and an efficient relevant feature selection method is proposed based on the 2D detection results and camera configurations.

The paper describes the process of generating dynamic object queries based on 2D detection results and camera configurations. It emphasizes the importance of focusing on specific objects and proposes an efficient method for associating bounding boxes in different views for a given object query. MV2D's performance is evaluated using different 2D detectors, and the results demonstrate its effectiveness in 3D object detection. The paper also includes a qualitative analysis to verify the localization capabilities of object queries generated from 2D detectors. MV2D achieves state-of-the-art performance on the nuScenes dataset, demonstrating its ability to recall objects and eliminate interference from noise and distractors. The framework's adaptability to different 2D detectors and its performance across varying input resolutions are also highlighted.

The paper discusses the effectiveness of relevant features, the impact of 2D detectors on 3D object detection, and provides a qualitative analysis of the framework's performance. It also addresses failure cases and potential limitations of MV2D. MV2D is effective in leveraging 2D object detection advancements for 3D object detection. The framework is able to generate more accurate 3D detection results by exploiting information from 2D object detection. However, one of the main limitations of this approach, is that it highly depends on the 2D detector, if the latter fails, MV2D will fail, too.

3 Project Structure

In this section, we present our project structure and progress till now. We will start by an overview of our Minimum Viable Product (MVP) and deliverables. Then, we will clarify the current project timeline, breaking down tasks and sub-tasks, assigning responsibility for each task, and providing reference dates.

3.1 Understanding the state-of-the-art

As our project revolves around research, it's crucial for us to dive deep into what's already out there—the state-of-the-art, as we call it. This means we spent a lot of time exploring complicated and condensed research papers, each looking deeply into a certain model. To make things more manageable, we've divided the task among ourselves. Each team member is responsible for reading a specific research paper, this way, we cover more ground efficiently.

3.2 Dataset and Data preprocessing

3.2.1 Dataset LIDC-IDRI

The LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) dataset stands as a cornerstone in medical imaging research, particularly in the domain of lung cancer detection. Comprising a rich collection of 1010 scans from 1018 patients, this dataset is a valuable resource for advancing our understanding and capabilities in medical image analysis.
The uniqueness of LIDC-IDRI lies in its representation of patient data through a series of slices, with each patient contributing over 100 slices. This dataset enables us to thoroughly investigate the complexities of lung structures, providing a detailed exploration of indicators like cancer cells.

3.2.2 Data preprocessing

The importance of the data preprocessing phase in enhancing the 3D localization from multiview X-ray images cannot be overstated. Our first step is to use the Digital Reconstruction Radiography (DRR) module to reconstruct lung images from the slice sizes provided by the LIDC-IDRI dataset. This reconstruction is necessary to ensure clarity on the later phase of our project, and we will convert the raw data into a format suitable for analysis.
Equally important are the extracted and used annotations made accessible by the `pylidc` library. These annotations, initially 3D, will be transformed into 2D coordinates in order to use them for training our Faster R-CNN for 2D detection. We will need the 3D annotations for the training phase. To simulate the multi-view scenario, we define images from 10 different angles in the pulmonary system presentation process.

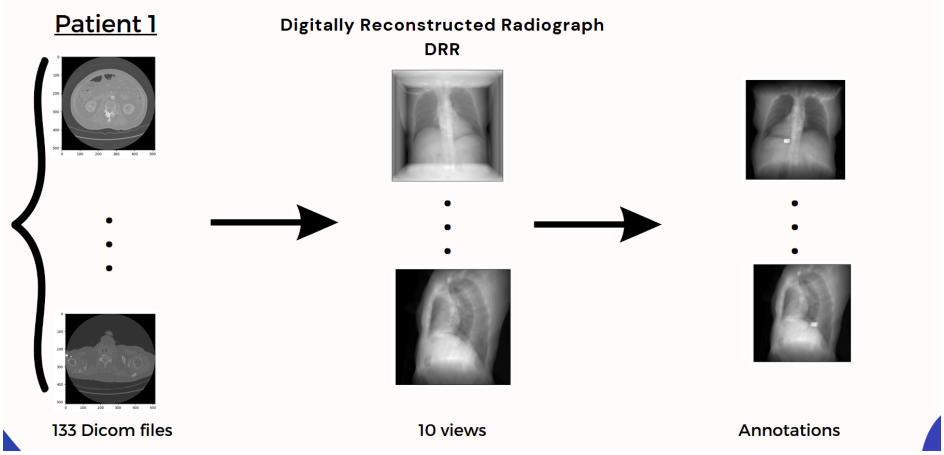

FIGURE 2 – Preprocessing

3.3 Model selection

3.3.1 First architecture

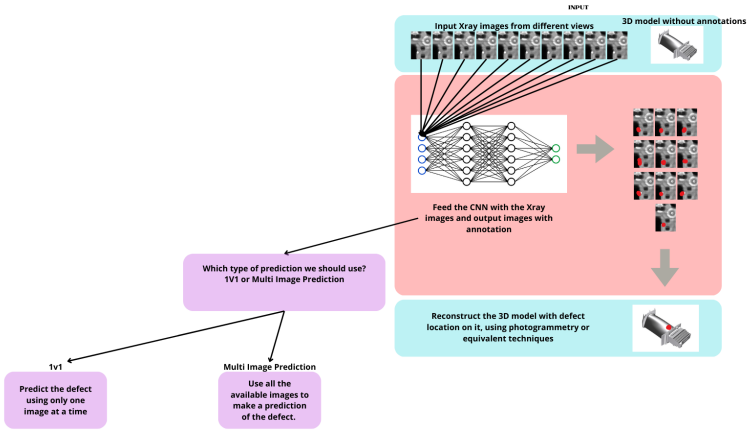

FIGURE 3 – First architecture

The cornerstone of our project lies in the development of an initial model architecture focused on defect detection within X-ray images. Drawing inspiration from successful implementations like LF-YOLO [1], we adopted the Deeplabv3 with MobileNet V2 as the backbone. This choice is informed by the algorithm's efficiency in handling X-ray image data and its proven success in identifying weld defects.

The model ingests a sequence of 10 X-ray images, processing each image individually to detect potential defects. The utilization of a convolutional neural network (CNN) allows the model to learn intricate features and patterns associated with defects. The LF-YOLO algorithm served as a benchmark, guiding our approach towards a faster and

lighter solution, optimizing computational resources without compromising accuracy [1].

During training, the model learns to recognize patterns indicative of defects, refining its understanding through backpropagation and gradient descent. The sequential processing of images aims to comprehensively analyze the entire object and identify defects present in each X-ray slice.

3.3.2 Photogrammetry and 3D Reconstruction

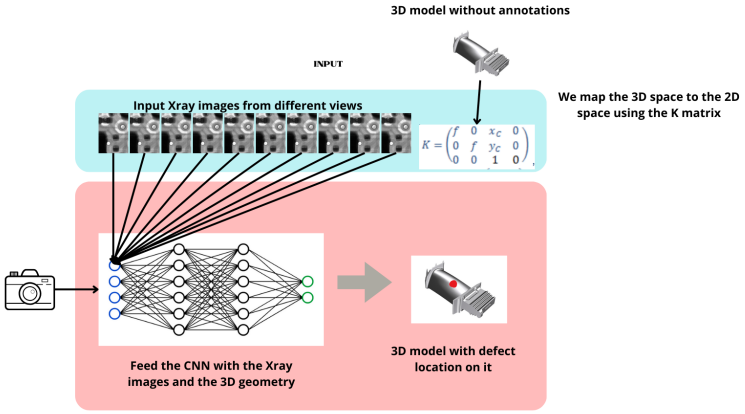

FIGURE 4 – Second architecture

The initial project phase involved the ambitious task of reconstructing the 3D geometry of the object using photogrammetry. However, challenges arose as the model lacked the capability to seamlessly integrate geometric information. This led to a pivotal redesign of the pipeline to accommodate both image analysis and geometric insights. To address this, the project incorporated a matrix representation of the object's geometry. This matrix includes critical parameters such as the focal length f and image coordinates of the projection P. The introduction of 3×3 rotation matrices R and 3×1 translation vectors T for each of the 10 views $V = 10$ views allowed encoding the point of view from which X-ray images were acquired.
This matrix-based representation facilitated the mapping of 3D space to 2D space through projection, providing the model with a nuanced understanding of the object's geometry in the context of X-ray images. The integration of this geometric information was pivotal in overcoming the limitations of the initial pipeline and laid the foundation for more accurate defect localization in subsequent phases.
The redesigned pipeline offers a novel advantage – simultaneous analysis of X-ray images by exploiting the geometry of acquisition. This concurrent processing enhances defect detection by considering complementary information from multiple perspectives. While the current approach relies on heuristic rules for matching detections in multi-

view 2D images, the proposed end-to-end approach aims to learn these heuristics directly from the model of the ideal part and the geometry of acquisition.

4 Task allocation

In our team organisation, we have divided our project in several parts : State-of-the-art, data preprocessing, Model selection, understanding the code, completing the code, training, testing and fine-tuning. For each part, we tried to see to what extend it was possible to assign different tasks to each one of the team, and how to assign tasks in the best way. For the initial task, *-state of the art-*, we naturally came to the conclusion that each of us could work on a different document. Thus, Massine worked on *Multi-view X-Ray-RCNN* and *Object as Query : Lifting any 2D Object Detector to 3D Detection*, Riccardo studied *Enhanced Defect Localization in X-Ray Images through Integrated Geometric Information, Training and fine tuning of Faster RCNN*, Thomas explored *ulti-view Vision Transformers for Object Detection*. Additionally, Khalil and João Pedro respectively worked on *A Light Touch Approach to Teaching Transformers Multi-view Geometry* and *Quad-DI¨P for X-ray cargo image*. This initial task, that gather working on research papers and on the five approachs on the image below, lasted from september to november.
For the other tasks, each one moves at his own pace but we, all make sure that we understood well our client's needs and that we are all on the same page. To do so, we often discussed the ideas that we have and confront them according to the client's requirements.
Following the *State-of-the-art*, other tasks emerged,such as analyse the datasets that we had to use and preprocess if necessary. These tasks were done by Khalid and Thomas for the analyse of the nuScene dataset and Massine for the preprocessing and the use of the LIDC-IDRI dataset. In the end, Khalil worked on testing the MV2D model with the Nuscene dataset, Riccardo was seeing how do fine-tuning and testing on the faster R-CNN model, while Massine was on the Geometry-based detection model.

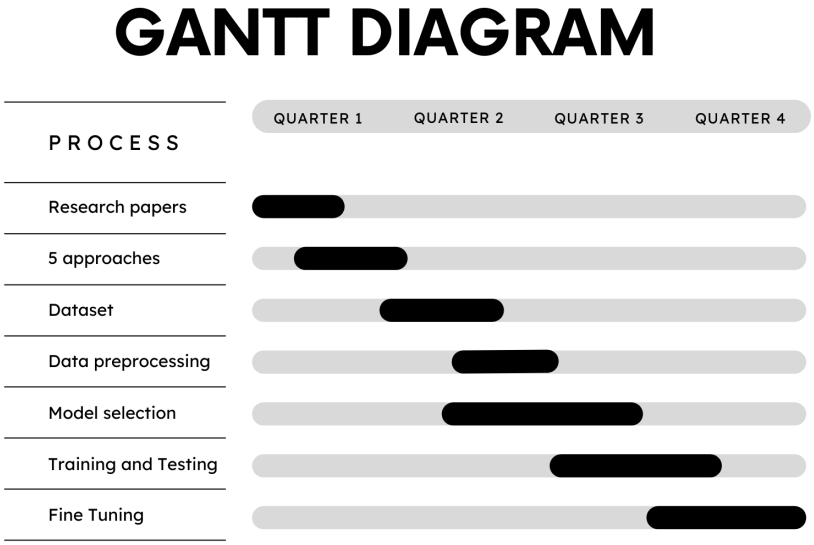

FIGURE 5 – Gantt diagram of our project

5 Technical elements & Deliverables

After our discussions with our clients, we decided to implement the following models :

— MV2D approach

— Faster R-CNN fine-tuning

— Geometry-based 3D detection

5.1 MV2D approach

The implementation of the MV2D framework in our project encountered several challenges, primarily stemming from the specificity of the model to particular datasets. The MV2D is designed to be compatible with datasets such as KITTI, nuScenes, Lyft, ScanNet, s3dis, SunRGBD, and Waymo. This specificity presents a double-edged sword : while it allows for fine-tuned performance on these datasets, it poses substantial barriers when attempting to extend the framework to alternative datasets. The highly specialized nature of the model's codebase, with functions intricately tailored to the original datasets, requires extensive refactoring for any adaptation to new datasets. This makes the process time-consuming and technically demanding, potentially hindering the project's scalability and adaptability to newer datasets.

The nuScenes dataset, which is approximately 512GB in size, exemplifies the difficulties in handling large datasets. The sheer volume of data necessitates a significant amount of storage and computational power, which can be a limiting factor for research teams with limited resources. Processing such large datasets requires a robust infrastructure, posing challenges not only in terms of computational resource requirements but also in data management and processing times.

Our project's progress was further impeded by dependencies on outdated software modules. The MV2D model implementation utilized versions of software modules no longer supported in our current computational cluster, leading to a range of compatibility issues. This lack of backward compatibility complicates the replication of the model's original environment, thereby affecting the project's reproducibility and the ability to leverage updates and improvements in newer software versions.

Setting up the correct environment for the MV2D model proved to be a complex task fraught with challenges. The environment issues ranged from conflicts between module dependencies to the difficulty in replicating the precise system configurations that the original model required. These setup challenges added a significant layer of complexity to the project, resulting in delays and additional resource allocation to troubleshoot and resolve these issues.

The codebase of MV2D, consisting of several nested functions specifically designed for a select few datasets, introduced additional complexity. The tight coupling between the code and these datasets made it challenging to integrate with other datasets. The framework's reliance on interdependent processes means that changing one part of the dataset requires a comprehensive understanding of the entire system to ensure that subsequent processes are not adversely affected. This interdependence limits the

framework's modularity and flexibility, creating a substantial barrier to extending its application beyond the initially intended datasets.

5.2 Faster R-CNN fine-tuning

This study employs the Faster R-CNN (Region-based Convolutional Neural Network) architecture to address the task of lung tumor detection, leveraging the LIDC-IDRI dataset (Lung Image Database Consortium and Image Database Resource Initiative). Faster R-CNN, proposed by Shaoqing Ren et al. in 2015, is a state-of-the-art two-stage object detection model that combines the strengths of deep convolutional neural networks (CNNs) and region proposal networks (RPNs)

The architecture of Faster R-CNN consists of three main components : a backbone CNN, a region proposal network (RPN), and a region-based CNN. The backbone CNN, typically pre-trained with well-established networks such as ResNet or VGG16, serves as a feature extractor. For lung tumor detection, this backbone processes input lung images to produce feature maps that capture hierarchical representations of image content.

The region proposal network (RPN) operates on these feature maps, generating bounding box proposals that are likely to contain tumors. It achieves this by sliding a small network, typically a 3x3 window, over the feature map and predicting multiple region proposals at each spatial location. These proposals are then ranked based on their objectness scores, and the top candidates are selected for further processing.

The selected proposals are then fed into the region-based CNN for classification and bounding box refinement. This region-based CNN utilizes a RoI (Region of Interest) pooling layer to align the proposals to a fixed size, followed by fully connected layers for classification and bounding box regression. During training, the model learns to classify the proposals into tumor or non-tumor classes and refines the bounding box coordinates to improve localization accuracy.

For training the Faster R-CNN model for lung tumor detection, the LIDC-IDRI dataset is utilized. This dataset contains a diverse set of annotated lung images, where each image is labeled with the locations of tumors in the form of bounding boxes.

5.3 Geometry-based 3D detection

The paper "3D Bounding Box Estimation Using Deep Learning and Geometry" proposes a method for estimating the 3D pose and dimensions of an object from a 2D image using deep learning and geometric constraints. The proposed method consists of three main steps : orientation estimation using a novel MultiBin module, dimension estimation using a fully connected network, and geometric constraint application using projective geometry. The results on Kitti dataset show that the proposed method outperforms state-of-the-art methods on all 3D metrics and achieves state-of-the-art results on the Pascal3D+ dataset. The proposed method does not require any 3D object models, making it more flexible and applicable to scenarios where 3D models may not be available or may be challenging to obtain.

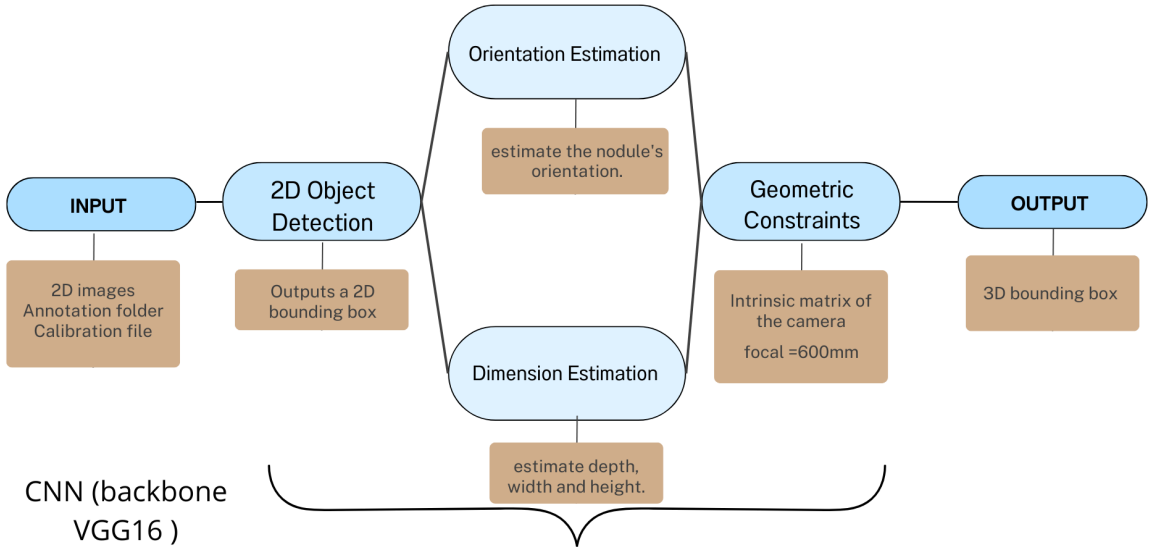

FIGURE 6 – Pipeline description

We implemented this approach using Tensorflow. We created our own custom dataset associated to LIDC-IDRI dataset to be able to apply our code directly on our dataset. One problem we encountered is that the shadow projection of the 3D annotations were sometimes misleading since the projections are from different angles. To alleviate that, we dilated the dimensions in the two directions (x, y) by 10 pixels. You can find this part's code here.

5.4 Performance evaluation

Performance evaluation in object detection involves measuring how accurately the system can detect objects in an image, and how well it is able to localize those objects.

There are several different metrics that can be used to evaluate the performance of an object detection system. Some common metrics include : Intersection over Union (IoU) : It measures the overlap between the predicted bounding box and the ground truth bounding box. It is calculated as the area of overlap divided by the area of union.

$$\textbf{IoU}(\mathrm{A}, \mathrm{B}) = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}}$$

A higher IoU indicates a better match between the predicted and ground truth bounding boxes. Furthermore, we use Precision which measures the percentage of predictions made by the model that are correct :

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

5.5 Training and Validation

5.5.1 Faster R-CNN fine-tuning

The training phase required several days of work, we were not able to download the dataset directly on the DCE cluster, so we decided to download 1/10 of the total data on our local machines, after some image preprocessing we uploaded the partitioned data on the DCE server and on Google Drive. To speed up the process we trained in parallel the model both on DCE and Google Colab. We tested several hyperparameters, and we performed training on different number of epochs. In the table below are reported the different combinations of hyperparameters tested :

TABLE 1 – Set of Tested Hyperparameters for Faster R-CNN

Optimizer	Learning Rate	Weight Decay	Number of Epochs	Batch Size
Adam	0.001	0.0001	50	16
SGD	0.001	0.0005	50	32
Adam	0.0005	0.0001	30	16
SGD	0.0005	0.0005	30	32
Adam	0.002	0.0001	40	16

After performing this tests we found that the best hyperparameters were :

TABLE 2 – Best Hyperparameters for Faster R-CNN

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.01
Weight Decay	0.0005

In evaluating the performance of our trained Faster R-CNN model for lung tumor detection, we employed a comprehensive set of results metrics, with a particular emphasis on Mean Intersection over Union (IoU). IoU, also known as the Jaccard Index, measures the degree of overlap between the predicted bounding boxes and the ground truth. It is calculated as the ratio of the intersection area to the union area of the bounding boxes. A higher IoU indicates better localization accuracy.

The Mean IoU is obtained by averaging the IoU scores across all instances in the test dataset. This metric provides a holistic measure of the model's ability to precisely delineate the identified tumors. A Mean IoU close to 1.0 signifies excellent alignment between predicted and ground truth bounding boxes, while lower values indicate reduced accuracy in localization. In our experimentation, we sought to optimize hyperparameters and model architecture to achieve a higher Mean IoU, reflecting improved segmentation performance and finer localization of lung tumors.

The presented Mean IoU (mIoU) results showcase a positive trend with increasing training epochs, implying an enhanced ability of our Faster R-CNN model to accurately localize lung tumors. Specifically, as the number of epochs progresses from 25 to

TABLE 3 – Mean IoU Results on Training (TR) and Testing (TS) Sets at Different Epochs

Epochs	mIoU (TR)	mIoU (TS)
25	0.262	0.267
30	0.286	0.288
50	0.311	0.308

50, we observe a steady increase in both mIoU on the training set (TR) and the testing set (TS). This trend suggests that further training epochs might yield continued improvements in model performance.

However, it is essential to note that our experimentation was constrained by a limited timeframe and a lack of available hardware resources. Given the observed positive correlation between mIoU and the number of epochs, it is advisable to explore even higher epoch values to potentially achieve a more refined and accurate model. Future investigations with extended training durations may provide deeper insights into the model's convergence and its ultimate ability to generalize to diverse lung images. Despite these constraints, the current results serve as a promising indication of the model's capacity to improve with additional training epochs.

The code is available at this link : `https://github.com/zari19/Faster-RCNN-fine-tuning`.

5.5.2 Geometry-based 3D detection

Fast training on the DCE proved to be incredibly beneficial in our project, significantly reducing the time spent on model training. To ensure the generalization capability of our model and prevent overfitting, we implemented an early stopping mechanism during the training process. The training process was set to stop when the variation of the loss function dropped below 0.001, and this occurred at the 25th epoch. Furthermore, to gain insights into the training dynamics and monitor key metrics, we used Tensorboard to get some graphs providing the evolution of our metrics and an intuitive overview of our model's performance throughout the training sessions.

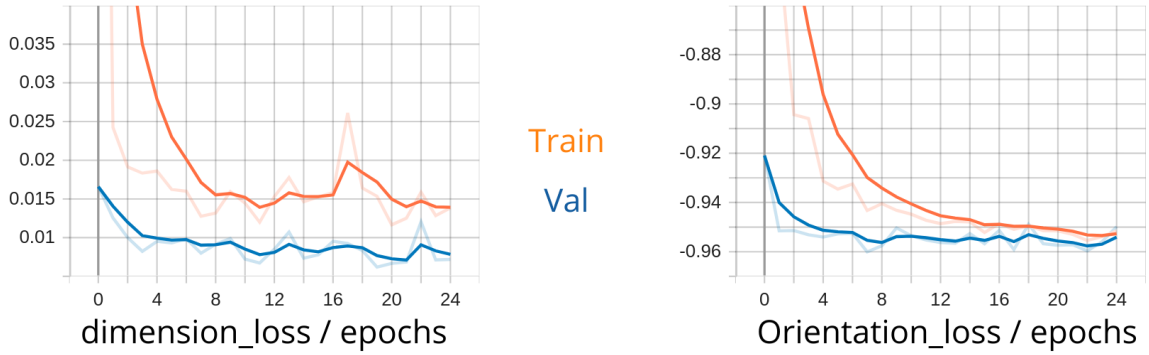

FIGURE 7 – Training and validation loss visualization

The training process demonstrated promising outcomes as the loss consistently

decreased for both the training and validation sets. This trend signifies effective learning and model convergence, indicating that our neural network is successfully capturing patterns within the data. Furthermore, our model achieved **92.2% precision** on the training set, showcasing its ability to make accurate predictions with a high level of confidence.

6 Conclusion

This report maps out our exploration into the complex world of 3D object localization from multi-view X-ray projections, a task that sits at the forefront of artificial intelligence and computer vision. Our work aligns with SAFRAN's innovative drive to push the boundaries of aerospace technology.

Our team has conducted a deep dive into the most advanced research available, critically examining a variety of techniques. The core of our project was the thorough analysis of the LIDC and nuScene datasets, which provided us with the rich data needed to train and validate our models. The detailed nature of these datasets has been invaluable, allowing us to refine our techniques in detecting lung abnormalities and enhancing autonomous driving technologies.

Despite encountering unresolved issues with the MV2D approach, our team persisted, demonstrating a deep commitment to advancing 3D detection technologies. We shifted our focus to fine-tuning the Faster R-CNN model, a process that demanded a nuanced understanding of lung imaging and the creative integration of geometric data for 3D detection.

We relied on the Intersection over Union (IoU) metric to evaluate our models' performance, a quantitative measure critical for ensuring the precision and reliability of our models.

The essence of this report is to document our journey—our strategic planning with SAFRAN, our analytical review of six state-of-the-art models, our selection and implementation of the chosen models, and the outcomes of our rigorous training and validation processes.

As we look forward, we recognize that the journey to refine 3D localization techniques from multi-view X-ray projections is far from complete. We are optimistic that our foundational work will inspire and guide future students working on this project. We are confident that the steps we have taken will contribute significantly to future advancements.

Références

[1] John Flynn Jana Kosecká Arsalan Mousavian, Dragomir Anguelov. 3d bounding box estimation using deep learning and geometry. *arXiv :1612.00496v2 [cs.CV] 10 Apr 2017*, 2017.

[2] Toby P. Breckon Brian K. S. Isaac-Medina, Chris G. Willcocks. Multi-view vision transformers for object detection.

[3] Faraz Saeedan Jan-Martin O. Steitz and Stefan Roth. Multi-view x-ray.

[4] M. Liu, Y. Chen, L. He, Y. Zhang, and J. Xie. Lf-yolo : A lighter and faster yolo for weld defect detection of x-ray image. 2021.

[5] Zitian Wang, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Object as query : Equipping any 2d object detector with 3d detection ability. *arXiv preprint arXiv :2301.02364*, 2023.

[6] Andrew Zisserman Yash Bhalgat, Joao F. Henriques. A light touch approach to teaching transformers multi-view geometry. *arXiv :2211.15107v2 [cs.CV]*, 2023.

[7] Gang Fu Yuxiang Xing Li Zhang Zheng Hu, Qiang Li. Quad-dip for x-ray cargo image decomposition.