SimPS-Net: Simultaneous Pose & Segmentation Network of Surgical Tools

Spyridon Souipas[1], Anh Nguyen[2], Stephen G. Laws[1], Brian L. Davies[1], Ferdinando Rodriguez y Baena[1]

***Abstract*—Localisation of surgical tools during operation is of paramount importance in the context of robotic assisted surgery. 3D pose estimation can be utilised to explore the interaction of tools with registered tissue and improve the motion planning of robotic platforms, thus avoiding potential collisions with external agents.**

With the problems of traditional tracking systems being cost and the need to redesign surgical tools to accommodate markers, there has been a shift towards image-based, markerless tracking techniques. This study introduces a network capable of detecting and localising tools in 3D using a monocular setup.

For training and validation, a novel dataset, 3dStool, was produced, and the network was trained to obtain a mean Dice coefficient of 85.0% for detection, along with a mean position and orientation error of 5.5mm and 3.3° respectively. The presented method is significantly more versatile than various state of the art solutions, as it requires no prior knowledge regarding the 3D structure of the tracked tools. The results were compared to standard pose estimation networks using the same dataset and demonstrated lower errors along most metrics. In addition, the generalisation capabilities of the proposed network were explored by performing inference on a previously unseen pair of scissors.

***Index Terms*—Surgical Tool Detection, Instance Segmentation, 3D Pose Estimation, Monocular, Surgical Tool Localisation**

I. Introduction

Throughout the past 25 years, Minimally Invasive Surgery (MIS) has been the focus of numerous developing technologies, and thus the number of Robot Assisted Surgeries (RAS) in this context has been constantly growing [1], not only due to the high precision and dexterity achieved, but also due to the potential efficiency gains offered [2]. Various surgical areas have received attention, including laparoscopic surgeries, orthopaedics and retinal surgeries. With the constant improvement of surgical robotics, the need to track tools involved in the procedure has been further underlined.

Detection of surgical tools has received significant attention, especially with recent improvements in computational resources and the development of deep learning techniques. A distinction can be made between image-based and non-image-based approaches for detection and 3D localisation. Non-image-based methods encompass either external sensors or mechanical solutions. The former usually manifests in the form of highly accurate, expensive and bulky optical trackers that are capable of tracking specific markers in 3D space. Although very effective in locating markers and straight forward in their use, such methods suffer from two major drawbacks, namely cost and the need to redesign surgical tools to facilitate the tracked markers. The latter drawback is especially significant when tracking ”off-the-shelf” tools such as scalpels and scissors, since the incorporation of markers would involve redesigning of the tools and manufacturing at high tolerance, thus further increasing cost per operation. Mechanical solutions, on the other hand, incorporate components such as motor encoders and cable driven components, which is the case with the da Vinci [3], to localise the tool-tip during operation. Such solutions also boast high accuracy, but require the development of a precisely manufactured system, which can be expensive.

With the emergence of computer vision practises in the past decades, detecting surgical tools within the operating room (OR) using image-based techniques has received significant attention. Deep learning, image-based techniques offer a desirable alternative to their non-image-based counterpart, since they significantly reduce cost and size of detection components. In addition, the tracked tools usually require no design readjustements.

Unfortunately, there exist several restrictions which prevent the direct application of image-based approaches in a surgical context to achieve tool detection and pose estimation. In endoscopic procedures, for example, tools are significantly close to the endoscope camera lens, which in turn amplifies the effects of image distortions. In addition, most surgical tools comprise of highly reflective or featureless materials, thus impeding the ability of a network to detect keypoints. Blood, smoke and other image occlusions present another obstacle in this process. Furthermore, several standard tools, such as scalpels and scissors are not accompanied by 3D models, usually in the form of Computer Aided Design (CAD) data, that can be used not only for point correspondences in detection, but also for pose estimation purposes. Finally, in order to estimate the pose of a tool in 3D space using cameras poses a significant challenge, especially with a monocular setup, which lacks any scene depth information. It should also be noted that, despite the high versatility offered by image-based solutions, they still underperform in terms of accuracy compared to the more expensive solutions offered by optical trackers.

With these limitations in mind, this paper introduces SimPS-Net, Simultaneous Pose and Segmentation Network that utilises a single RGB camera to segment “off-the-shelf” surgical tools within a frame and estimate their 3D pose in space. In addition, a novel dataset constructed for the purpose of training and testing this network is introduced.

[1] Mechatronics in Medicine, Imperial College London, UK ss8413@imperial.ac.uk

[2] Department of Computer Science, University of Liverpool, UK

This paper is structured as follows. Section II provides an investigation of relevant research on the topic of tool detection and pose estimation in a surgical context, along with results that verify the suitability of such methods. Section III outlines the dataset that was constructed for the purposes of training and testing the network. Section IV then gives an overview of the developed network, and the experimental setup used to test it, with the results being analysed in section V. Section VI utilises these results to compare the performance of the examined technique with respect to the state of the art, and in section VII an outline of future improvements is provided.

II. Related Work

With the improvement of computing resources and the development of Deep Learning (DL) techniques in computer vision, mostly in the form of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significant recent research has focused on developing tool detection pipelines that utilise neural networks, as opposed to previous, "traditional" techniques which usually perform feature extraction and matching or incorporate various sensors. This is not only because DL techniques usually demonstrate higher accuracies, but also because of their deployment convenience, assuming a labelled dataset is available. An extensive analysis of traditional methods for surgical tool detection has been undertaken [4], and these generally suffer from low accuracy in the case of camera based methods, or the need to redesign surgical tools in the case of optical trackers. For this reason, the use of standard RGB cameras is favoured, alongside a steep development of Deep Learning techniques for tool detection and pose estimation related research.

A. Deep Learning Detection

To achieve pose estimation using deep learning techniques, object detection in an image is usually a prerequisite. Therefore, when discussing pose, it is only natural to first examine various methods of surgical tools detection and segmentation. The two standard tool classification methods are box detection, achieved by creating a bounding box around each tool in the frame and classifying the contents of each box, and instance segmentation, where pixel-wise classification is performed to identify what class each pixel belongs to. Note that when examining single object frames, the term semantic segmentation is used, which essentially identifies the regions of an image that belong to a tool, without specifying which tool. Several approaches have been developed to achieve bounding box tool classification. One interesting implementation examined the performance of a You Only Look Once (YOLO) based network [5] across 7 different types of tools [6]. YOLO based networks boast very high performance speeds, making them an ideal choice for real time applications. However, with the lack of optimisation, they underperform in harsh environments such as the ones associated to laparoscopic or orthopaedic operations. To address this, a YOLO900 based network [7] was constructed to take into account motion prediction from previous frames in order to improve detection performance [8]. This use of temporal information, however, can lead to an exponentially increasing error, since the success of tool detection in the current frame depends on the success of previous detections. Nevertheless, this approach further underlines the high speed and tunability of YOLO based networks. These networks, however, are not the only ones that demonstrate high detection speeds. It has been shown that it is possible to achieve extremely fast box detection by constructing non-region-based networks for tool detection. Specifically, the Extremely Fast and Precise Network (EF-PNet) [9] was constructed with the aim of optimising detection speed, achieving 270 fps in detection. This research suggests that in a surgical context, YOLO based networks may not be optimal for box detection. A similar process was explored when developing a box detection network that did not require the formulation of anchors [10]. Even though the inference speed only ran at 37 fps, the accuracy was increased significantly. Another advantage of deep learning approaches was the option to integrate spatio-temporal data across the detection process to improve results. A Spatial Transformer Network (STN) has been combined with a CNN to detect tools moving at high speed, a condition which usually suffers from erroneous detections due to image blur [11]. However, occlusions significantly impact such methods.

With the development of U-Net [12], numerous techniques for surgical tool segmentation have been proposed to address the difficulties presented in the operating room. Furthermore, detection could now be effectively performed on non-rigid tools. It was shown that with minimal modification, a standard Fully Convoluted Network (FCN), namely the FCN-8s [13], was adjusted to semantically segment surgical tools in operation. The results of this network were further combined with the detected optical flow across subsequent frames to further refine the segmentation [14]. Expanding on the FCN-8s network, two variations of a new network, ToolNet [15] were developed for real-time segmentation, namely ToolNetMS and ToolNetH. Both ToolNet versions benefit from fewer network parameters than standard FCNs. ToolNetMS achieves faster inference at 43 fps, whereas ToolNetH boasts a higher segmentation accuracy. A different approach explored for segmentation was the construction of a hybrid structure, consisting of both CNN and RNN elements [16]. The convolutional layers extract and encode features across image pixels, whereas the recurrent layers identify pixel dependencies. The hybrid structure was a significant improvement from previous implementations.

The effect of data augmentation in segmentation results has also been explored to an extent. Initial segmentation state of the art networks such as U-Net and TernausNet directly used image masks for training. However, surgical environments significantly vary across operations, and therefore augmented data could potentially account for this discrepancy. UNetPlus, a network that brought together aspects of both U-Net and TernausNet was therefore constructed [17], and a novel augmentation technique was employed throughout training, proving that detection results greatly benefit from image augmentation. The effect of augmentation is further underlined when comparing

UNetPlus to other U-Net - TernausNet hybrids which do not incorporate data augmentation [18]. Instance segmentation can be taken one step further by defining different classes across a single object, such as for example the tool shaft and blade on a scalpel. This was first achieved by employing a standard Residual Network, ResNet-101, and training it to detect various parts of each tool [19]. Besides the accuracy improvements that this network benefits from, it has also proven successful for occlusion handling, since segmentation of an object will not fail completely if a region of the tool is occluded. In this case, occlusions over the tool shaft still allow for tool tip segmentation and vice versa. Segmentation networks also incorporated temporal information through the construction of three dimensional convolutional layers. For example, laparoscopic videos have been used across an encoder-decoder architecture that includes three dimensional convolutional layers to improve segmentation results [20] and identify tool landmarks. Such architectures address detection errors caused by unfavourable background conditions.

B. Pose Estimation

Significant effort has been directed towards extracting the pose of an object through a monocular RGB camera using deep learning techniques. Early approaches were not, however, applied on surgical tools. The majority of the solutions approach the estimation of object pose either as a regression problem, a template matching problem or, more rarely, via a coordinates based approach [21]. Usually, two-stage methods are employed, which involve a network that achieves object detection, followed by a pose estimation pipeline. Regression methods were initially applied on large object pose estimation with the development of PoseNet [22], however networks were further refined to address everyday objects. An example is the PoseCNN [23], where an object is localised in an image using a bounding box, and the 6D posed is regressed within the detected region. In template matching, the pose is solved through a PnP problem using correspondences between image points and 3D model points. Correspondence methods are more frequent, but require 3D information of the detected object, usually extracted through CAD models. These methods sometimes boast increased robustness against occlusion by matching numerous pixels against the 3D model [24], or fast running times, as is the case of EfficientPose [25], which can localise objects and estimate the respective pose values in real time. Another noteworthy pose estimation technique, which was a direct result of the development of the Single-Shot Detection (SSD) [26], is the SSD-6D [27]. In this case, objects are detected in a single RGB frame using the SSD network in the form of a bounding box, and the 6D pose is subsequently estimated via viewpoint classification instead of translation and rotation regression.

Several approaches have been developed to detect tools and estimate their pose within images. Most of these are two stage approaches, though some employ a single-step structure where detection and pose estimation are achieved within the same network. Interestingly, a lot of effort has been directed towards 2D pose estimation using monocular images instead of 3D pose, as it is easier to label data for network training in this manner. One initial single-stage implementation is the direct regression of 2D tool pose in an image within an encoder-decoder network [28]. In such a case, tool joints are identified within an image alongside tool segmentation. This process is then further refined to localise these same tool joints by obtaining heatmaps for each landmark and identifying small regions that contain the points of interest, which are then overlayed to the original image. A similar approach was developed to segment tools using a U-Net and simultaneously localise tool landmarks in an image through the use of probability maps [29]. These two approaches were the first single-stage techniques to be applied on surgical tools to not only detect them, but also estimate their 2D pose. In theory, if the definition of 2D pose is reduced to identifying tool landmarks, a detection network alone could be sufficient for this task. However, when comparing such an approach to a two-stage solution, the benefits of the latter become clear. The standard semantic segmentation networks, which are usually encoder-decoder types, usually only offer dense pixel segmentation. In one such example [30], a standard encoder-decoder network was initially used to estimate joint positions. However, it quickly became obvious that any network would benefit from some mechanism dedicated to pose estimation. In this example, a regression stage was applied on the dense segmentation results to extract accurate pixel location of tool joints. Similarly, attention methods instead of regression have been employed to further refine the pixel region containing tool landmarks in 2D pose detection [31].

Examining 3D surgical tool pose using monocular RGB cameras through deep learning is still relatively unexplored. Following the development of the SSD-6D [27], some effort was directed towards applying it on surgical tools [32]. To do so, it was shown that the structure of SSD-6D needed to be altered so that pose estimation could be tackled as a regression problem instead of a classification one. The network was trained on artificial images, and in doing so utilised the required 3D model information, but some degree of generalisation was presented when using real images. Pose estimation on artificial images was further refined in [33], where different rendering options were explored for the artificial dataset, and the pose model was adjusted accordingly. Lastly, a noteworthy approach was developed to address the need for 3D models in 3D pose estimation. In laparoscopic surgery, it is safe to assume that most tools observed by the camera will be cylindrical, and therefore the Augmented Reality Tool Network (ART-Net) [34] was developed to include a single encoder followed by multiple decoders. One of the decoders is used for detection, another for segmentation, and the remaining ones are used to extract tool edge-lines, mid-axis line and tool tip, from an image. These parameters are refferred to as geometric primitives, and knowledge of their 3D location can be used to solve simple algebraic statements that can provide the pose of cylindrical objects.

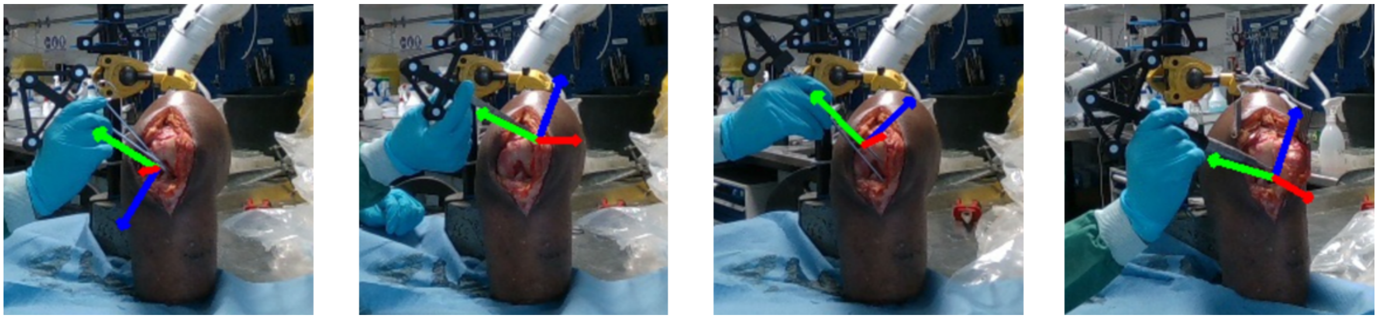

Fig. 1: Examples of true poses in 2D (Forceps, Scissors, Burr, Scalpel from left to right)

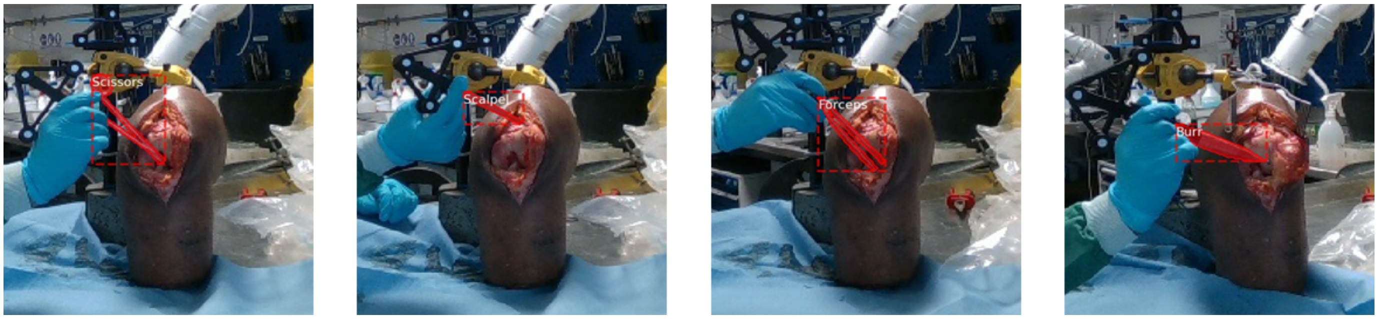

Fig. 2: Examples of manual image annotations (Scissors, Scalpel, Forceps, Burr from left to right)

III. Dataset

Most datasets of surgical tools available examine laparoscopic conditions and usually do not provide any labelled images of ”off-the-shelf” tools. In addition, such datasets usually lack 3D pose labels. Hence, a custom dataset was required to train SimPS-Net. Specifically, the dataset should consist of RGB images of surgical tools in action, alongside the 3D pose of each tool. For that purpose, the novel dataset, 3D Surgical Tools (3dStool) was constructed. Four surgical tools were initially chosen, namely a scalpel, a pair of scissors, a pair of forceps, and an electric burr. The first three of these are extremely common in surgical operations. An electric burr, while less common, was also included in order to explore the response of the network to axisymmetric, cylindrical objects. All objects were recorded while operating on a cadaveric knee* to best mimic the real-life environmental conditions.

Overall, 5370 images were collected and annotated, split into 4027 images for training, 537 for validation and 806 for testing purposes. Each image was semantically labelled and was accompanied by a value for the observed 3D pose. Additionally, while the main dataset consists of only one variation of each tool, a separate set of 553 images of a different set of scissors was collected, annotated, and used to explore the generalisation capabilities of the network. Ultimately, the presented dataset is publicly available, accompanied by a detailed explanation of the structure and some relevant functions for image processing [35].

IV. Materials and Methods

This section initially outlines the dataset construction, followed by the structure of the proposed network and, ultimately, the hardware employed.

A. Dataset Collection & Extrinsic Calibration

For each tool, two different sources of data were required. Firstly, the RGB image of the tool in action, which would subsequently be manually annotated; secondly, the 3D pose of the observed tool in the camera frame. For the purpose of RGB image collection, a RealSense D415 camera (Intel, USA) was employed. For the collection of position data, the ftk500 optical tracker (Atracsys, Switzerland) was utilised, allowing for localisation of fiducials in 3D space. The two sensors were rigidly positioned relative to each other, and extrinsically calibrated in order to obtain the 3D poses in camera coordinates, as outlined in Equations 1 and 2:

$$\begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_A \\ y_A \\ z_A \\ 1 \end{bmatrix} \tag{1}$$

$$\therefore \ \mathbf{p_{cam}} = [\mathbf{R}|\mathbf{t}] \ \mathbf{p_A} \tag{2}$$

where $[\mathbf{R}|\mathbf{t}]$ is the extrinsic calibration matrix, $\mathbf{p_A}$ are the 3D coordinates of a point in the optical tracker frame and $\mathbf{p_{cam}}$ are the same 3D coordinates of a point in the camera frame. Note that a similar transformation was undertaken to

*The knee was obtained from a licensed tissue bank, and ethical approval was acquired from the Imperial College Healthcare Tissue Bank, project code R21046.

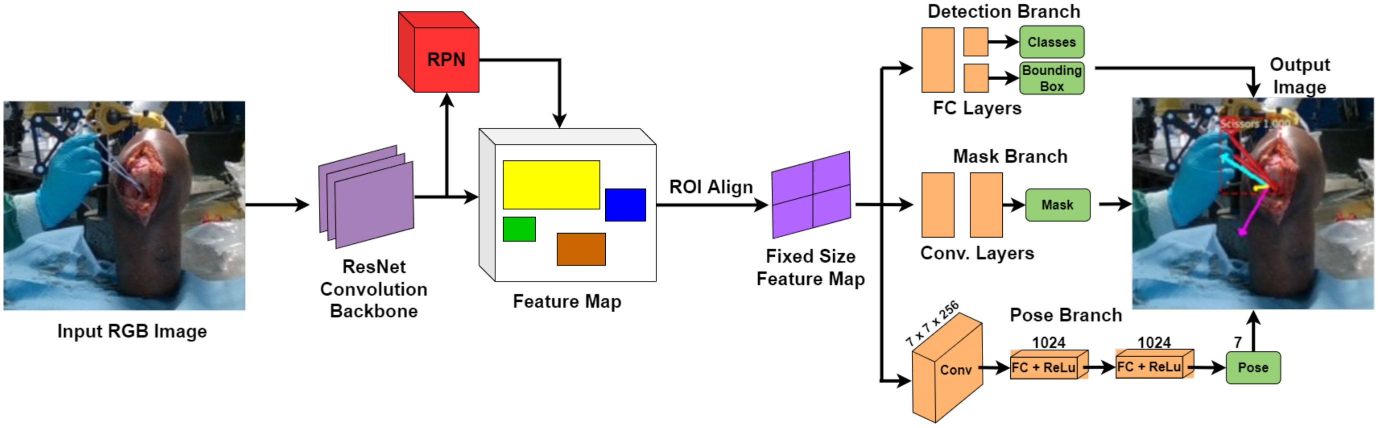

Fig. 3: SimPS-Net Architecture

calculate the orientation of each tool in camera coordinates. In order to estimate $[\mathbf{R}|\mathbf{t}]$, an implementation of the standard Iterative Closest Point (ICP) algorithm [36] was employed. Specifically, a geometry of known shape was designed and 3D printed, with fiducial markers also attached to it. The geometry was subsequently placed before the camera and the optical tracker. The RealSense D415 was used in depth mode to collect a pointcloud of the detected geometry, which was then matched to the CAD model to create an initial estimate of the transformation. ICP was then utilised to further refine the transformation matrix, allowing for the calculation of $[\mathbf{R}|\mathbf{t}]$.

Having obtained the extrinsic matrix, 3D printed clamps were utilised to incorporate optical tracker markers on each surgical tool. The tools were recorded in operation, with the background set up to best mimic actual operating room conditions. For this purpose, data was collected while operating on a cadaveric knee. Each batch of images was accompanied by the calculated 3D pose, which was subsequently converted to camera coordinates. For visualisation purposes, the intrinsic matrix of the camera, $\mathbf{K}$, was used to visualise the 3D pose using the 2D RGB images, as shown in Figure 1. It should be noted that some sources of error, such as slight vibrations of the sensors, or displacement of the tool clamps, could have caused some variation between the tracked and the actual position of the tool.

Upon verifying the pose data for each batch of images, the batches were combined, shuffled, and split. Images where no pose was detected due to heavy occlusion were rejected. Ultimately, polygon annotations were manually generated, as shown in Figure 2, for training and testing of the network. The poses were included in an annotation file in the form of translation (x, y, z) and orientation (4 quaternions).

B. The Network

The proposed network architecture allows for multi-instance segmentation and 3D pose estimation. It is based on the famous Region Based CNN, Mask-RCNN [37], which is widely used for semantic segmentation. Mask-RCNN comprises the backbone, which is followed by two branches, one responsible for segmentation and the other for classification. SimPS-Net expands on the Mask-RCNN by introducing a third branch which performs 3D pose regression, as shown in Figure 3. The novelty of the proposed approach is represented by the ”Pose Branch” in the figure, since the rest of the architecture is approximately similar to that of Mask-RCNN. As observed, the examined branch comprises of an initial convolution step, where the coloured image is passed through the convolutional layers. During this layer, a 7x7 pooling is applied to extract features from the original image being passed through the layer. Note that to maintain a consistent output dimension, no zero-padding was undertaken during the initial step. Upon extracting the necessary features during the convolutional layer, the output is then passed through two identical, fully connected layers. In doing so, the general features extracted during the first step of the branch are further refined, in order to better understand the position and orientation of the detected tools. Ultimately, the results are passed through a dense layer, which utilises a Rectified Linear Unit (ReLU) activation, thus expressing the output in the form of seven parameters, namely the position and orientation values. Various activation methods were explored, such as the sigmoid and the softmax methods, with ReLU achieving optimal results.

The network was set up using ResNet-50 as the backbone to extract the regions of interest, according to relevant findings that indicate improved surgical tool detection in doing so [38]. Minor alterations were made in the classification and mask branches, in order to allow for faster deployment.

In this application, the 3D pose, $\mathbf{p}$, is defined as the combination of the position and orientation vector for an identified object. Quaternions were used to describe the orientation of each tool:

$$\mathbf{p} = [\mathbf{x}, \quad \boldsymbol{\theta}]$$

where

$$\mathbf{x} = [\mathrm{x}, \quad \mathrm{y}, \quad \mathrm{z}] \text{ is the position vector}$$
$$\boldsymbol{\theta} = [\mathrm{q_x}, \quad \mathrm{q_y}, \quad \mathrm{q_z}, \quad \mathrm{q_w}] \text{ is the orientation vector}$$

In this context, the pose branch predicts the 3D pose of

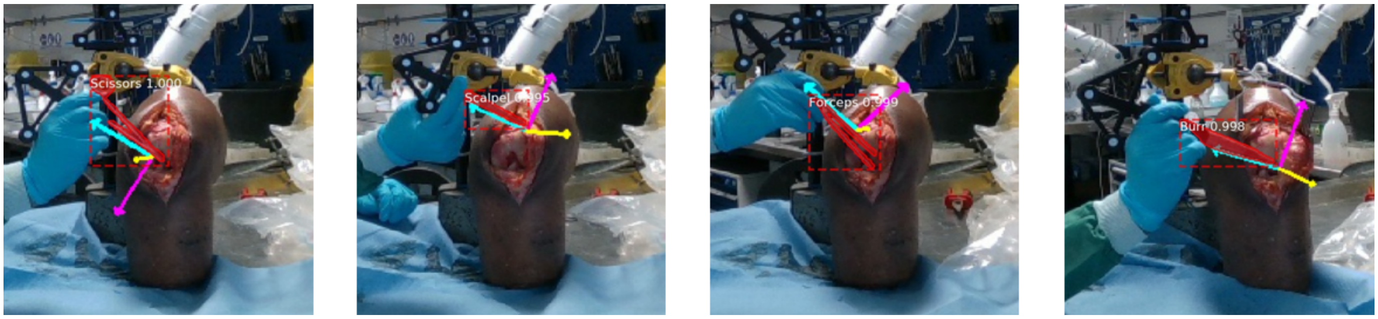

Fig. 4: Examples of predicted masks and poses in 2D (Forceps, Scissors, Burr, Scalpel from left to right)

each tool within the frame, $\mathbf{p_{pred}}$. The predicted result is then compared to the true 3D pose of the detected tools, $\mathbf{p_{true}}$. This comparison is undertaken in the final step of the branch, where the pose loss function is constructed to compare position and orientation separately, before the two are amalgamated to calculate the overall loss. This is demonstrated in Equation 3.

$$\mathcal{L} = \alpha \ \left\| \mathbf{x_{true}}, \mathbf{x_{pred}} \right\|_2 + \beta \ \left\| \boldsymbol{\theta_{true}}, \boldsymbol{\theta_{pred}} \right\|_2 \qquad (3)$$

The constant β has been proven to improve orientation prediction [22], whereas α is needed to account for the discrepancy in scale between the values of the orientation and the position vector, since position values were incorporated in meters.

C. Hardware

The network was trained and deployed using an NVIDIA GeForce GTX 1060. Training was undertaken for 160 epochs. The current implementation achieved inference at 2.2 fps.

V. Results

Evaluation of SimPS-Net was achieved by presenting unseen images of all four classes to the trained network and performing tool detection and pose estimation for each frame. This was repeated for various permutations of pose constants, α and β, to identify the best configuration of the network, by examining the errors along the position and orientation values obtained. Inference was also undertaken for a different pair of scissors to explore the extent to which SimPS-Net detection and pose estimation can be generalised across ”off-the-shelf” items.

A. Inference Results on Dataset

Detection, position and orientation metrics were obtained for each permutation across all frames. The pose constants themselves had minimal impact on the detection metrics, and therefore the detection findings reported in Table I remained relatively constant across tests. Examples of the predicted masks, along with the predicted poses, are shown in Figure 4. The metrics used for detection were the mean average precision (mAP) and the mean Dice coefficient (mDice). For pose estimation, the output was processed in mm and degrees.

To better explore the effect of pose constants, the network was initially trained so that position and orientation could be inspected separately. This was done by setting one of the pose constants to zero and considering various values for the non-zero constant.

Training was then undertaken with various constant permutations to converge on a suitable combination. When merging both components of the loss function, the better pair was found to be $\alpha = 700$, $\beta = 300$. The inference error for both position and orientation was ultimately calculated as the mean difference between predicted and true vectors across all test images along each axis. The results of this proposed method are reported in Table I.

B. Inference Results on Unseen Scissors

The same configuration of α and β was then used to perform inference on the pair of scissors that had not been included in the training process beforehand. The purpose of this test was to understand whether or not the network could easily generalise on surgical tools that have not been seen before, since tool shapes and sizes can deviate to an extent, depending on the manufacturer. The network was used to infer the poses of the

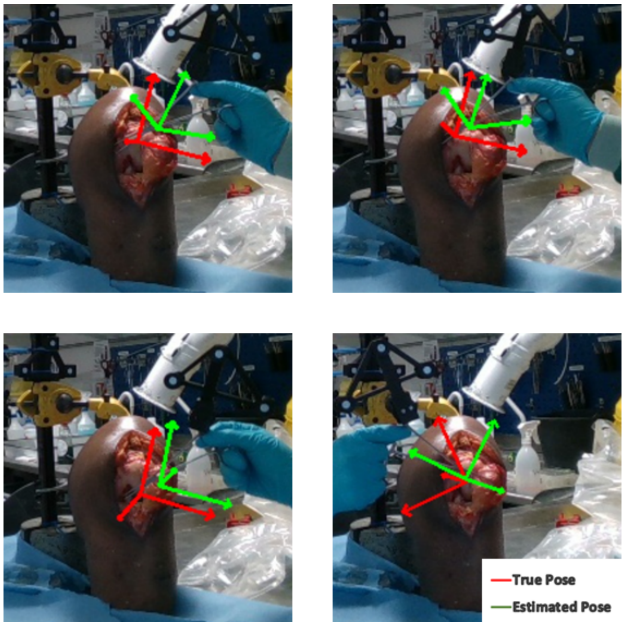

Fig. 5: Pose Estimation on Unseen Scissors

scissors, which are shown alongside the ground truth poses in Figure 5.

VI. Discussion

The main benefit of this network is improved versatility over competing designs. No prior knowledge of the 3D structure of the tools is required to achieve results that generally exceed the state-of-the-art in 3D pose estimation. Furthermore, no initial assumptions regarding the shape of the tools are required, as opposed to methods which assume the surgical tools can be approximated to, e.g. cylinders.

SimPS-Net was compared to three networks which allow for 3D pose estimation, specifically PoseNet [22], Robust Object Pose Estimation (ROPE) [39], and Geometry-Guided Direct Regression Network (GDR-Net) [40].

PoseNet allows for the estimation of camera position and orientation by regressing quaternions, without the need for any prior 3D structure knowledge. Specifically, it involves a CNN that can estimate camera pose over 6 degrees of freedom by utilising a single, RGB image. The purpose of developing this network was to address pose estimation cases of varying fields, both indoors and outdoors, whilst also accommodating severe cases where harsh lighting or other occlusions are present. The authors first deployed this network for the purpose of calculating camera pose when examining big objects, such as buildings, however it is expected that the network can be deployed in closer proximities

ROPE allows for some degree of occlusion handling in detection and 3D pose estimation. Interestingly, when exploring the architecture of this competing network, it is noted that the authors also utilised a Mask R-CNN as a basis, however the mask branch was also changed to allow for 2D image landmark identification. The landmarks are subsequently used to estimate the 6 degrees-of-freedom pose of the detected object on the scene. The landmarks are then matched to the object, and through a standard Perspective-n-Point (PnP) method [41], the pose of the object is calculated in 3D space.

Finally, GDR-Net makes use of correspondences using geometric representations. More analytically, the network initially achieves detection of all relevant objects in a single RGB image using a standard detection network. Subsequently, for each object, a zoomed in sub-section is extracted from the image, containing only the object of interest. This sub-section is utilised in the network to generate geometric feature maps. Ultimately, these feature maps are used in order to regress the 3D position and orientation of the target object.

As shown in Table I, the position and orientation metrics obtained for the proposed method are promising. Indeed, none of the position errors along each direction exceeds 10mm, as opposed to the state of the art results. In addition, orientation errors remain low throughout. Pitch and yaw are similar when compared to the other networks. Interestingly, SimPS-Net appears to be more robust along the roll axis than its counterparts, which, along with the results along the depth direction (Z), indicates suitability for 3D tasks. The reported errors can be amalgamated to a mean positional error of 5.5mm and a mean orientation error of 3.3°. As expected, these accuracies cannot outperform optical trackers, but the ease of deployment can counteract this shortcoming, making this the preferred method for applications that do not require submilimiter accuracies, such as robot path planning outside the body.

Another interesting finding is associated with the orientation estimation in axisymmetric tools. All four tools employed in this survey were almost identical on either face, with minimal feature discrepancies. Irrespective of this, the network managed to consistently identify the exposed tool side and correctly calculate the 3D pose in the majority of the cases in the constructed dataset.

Unfortunately, the detection results are worse when compared to competing networks which focus solely on detection. This, however, can also be mitigated by editing the Mask-RCNN branch of the network to address cases of occlusion, which are frequent in the context of surgical operations. As previously discussed, operating rooms suffer from a plethora of detection impediments, some of which have been extensively addressed in other pieces of research and could be integrated in the proposed network in the future.

Some degree of generalisation has also been achieved by the network. Figure 5 demonstrates the differences between true and estimated pose in four instances of a previously unseen pair of scissors. Even though the positional errors have increased compared to the larger dataset, it is important to note that, to some degree, localisation is successful. A trade-off is noted between orientation accuracy and positional accuracy. For example, the top two images demonstrate relatively low orientation error, with more significant positional errors, whilst the bottom right image achieves low positional error, but at the cost of higher orientation error. The bottom left image also demonstrates a low orientation error, but it is interesting to note that in this case, the network has failed to identify the correct face of the pair of scissors, and therefore one of the axes has been offset by 180°. These findings suggest that the chosen permutation of α and β may need to be further optimised.

Not many techniques utilise a monocular setup to estimate the pose of surgical tools in 3 dimensions without fusing any other information. In fact, as stated in Section II, most research has focused on 2D pose, which reports position error in terms of pixels. While useful, such methods do not provide 3D localisation of tools in the OR, and therefore any results comparison would provide limited insights.

Despite the encouraging results presented in this paper, however, the generalisation capabilities of the proposed network are still obstructed by the limitations of the dataset in its current state. Specifically, the use of a single cadaveric knee for data collection does not allow the network to be optimised for the case of different patient skin colours. In practise, skin colour will be a significantly varying parameter, and therefore the dataset should be expanded to accommodate this concern. Additionally, even though the image background was set up to mimic the conditions of an operating room, some annotated

TABLE I: SimPS-Net Results Comparison against Literature

Source	mAP (%)	mDice (%)	X (mm)	Y (mm)	Z (mm)	Pitch (deg)	Yaw (deg)	Roll (deg)
PoseNet [22]	NA	NA	18.4 (11.6)	18.6 (13.1)	13.4 (9.2)	2.3 (1.8)	1.3 (1.0)	28.2 (28.2)
ROPE [39]	56.8	80.2	8.4 (3.5)	11.4 (6.2)	9.2 (5.2)	3.2 (2.5)	1.8 (2.1)	8.3 (16.2)
GDR-Net [40]	58.5	83.7	6.1 (5.2)	5.0 (2.4)	7.3 (3.4)	2.6 (3.1)	2.3 (2.6)	6.7 (10.7)
SimPSNet	62.9	85.0	5.2 (4.5)	4.0 (4.3)	6.3 (6.0)	2.4 (2.8)	1.5 (1.5)	6.1 (37.3)

images of tools in actual operation would be very beneficial and further improve the network.

Moreover, considering the versatility of this network, endoscopic applications could be achieved. Nevertheless, a new dataset should be generated for the purposes of training, which should comprise of endoscopic tools in action, with manual annotations and 3D pose accompanying the collected images. Since the use of optical trackers for pose estimation is not an option for motion of fully occluded tools, an alternative technique, such as robot kinematics, should be utilised instead. However, as discussed in Section II-B, endoscopic tools can be represented as cylindrical objects, and therefore there exist networks that are more appropriate for this purpose, such as the ART-Net [34].

Regardless of the aforementioned limitations of the dataset, this paper introduces an inexpensive, monocular camera-based method capable of not only detecting, but also localising standard surgical tools in operation in 3D space. SimPS-Net requires no prior information regarding the shape of the tools, unlike the majority of pose estimation networks. Furthermore, when compared to similar state of the art methods, the proposed technique managed to outperform other networks across all six examined degrees of freedom.

VII. Conclusion and Future Work

This study has established a successful methodology for simultaneous segmentation and 3D localisation of surgical tools upon inspecting a single RGB camera image. A new dataset was created that incorporates manually labelled masks along with the associated tool 3D position and orientation in the form of quaternions. A Mask-RCNN has been modified to regress the 3D pose of detected tools. The errors obtained using the proposed network along with the constructed dataset outperform other networks that allow for 3D pose estimation without any prior 3D tool structure knowledge, in both position and orientation.

However, some areas need to be further surveyed. Specifically, the hardware available at the time of collecting the dataset did not allow for images that included multiple tools. Therefore, the pose errors will need to be explored when multiple tools are present in a frame. Additionally, the dataset should be expanded to include cadaveric knees of different skin colours. Ultimately, with an improved GPU, the real-time capabilities of the network should also be explored. Such a feat could lead the development of a visual active constraint that would allow for the integration of the proposed network in the worklflow of a surgical robot pipeline, thus making operation smoother and ensuring patient safety, whilst avoiding robot damage.

References

[1] Jocelyne Troccaz, Giulio Dagnino, and Guang-Zhong Yang. Frontiers of Medical Robotics: From Concept to Systems to Clinical Translation. *Annual Review of Biomedical Engineering*, 21(1):193–218, 2019.

[2] Daniel J. Lee, James Ding, and Thomas J. Guzzo. Improving Operating Room Efficiency. *Current Urology Reports*, 20(6), 2019.

[3] Michael E Moran. Epochs in Endourology The da Vinci Robot. *Journal of Endourology*, 20(12), 2006.

[4] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting Surgical Tools by Modelling Local Appearance and Global Shape. *IEEE Transactions on Medical Imaging*, 34(12):2603–2617, 2015.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016.

[6] Bareum Choi, Kyungmin Jo, Songe Choi, and Jaesoon Choi. Surgical-Tools Detection Based On Convolutional Neural Network In Laparoscopic Robot-Assisted Surgery. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 1756–1759, 2017.

[7] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, 2017.

[8] Kyungmin Jo, Yuna Choi, Jaesoon Choi, and Jong Woo Chung. Robust Real-Time Detection of Laparoscopic Instruments in Robot Surgery Using Convolutional Neural Networks with Motion Vector Prediction. *Applied Sciences*, 9(14):2865, 2019.

[9] Dongqing Zang, Gui Bin Bian, Yunlai Wang, and Zhen Li. *An Extremely Fast and Precise Convolutional Neural Network for Recognition and Localization of Cataract Surgical Tools*, volume 11768 LNCS. Springer International Publishing, 2019.

[10] Yuying Liu, Zijian Zhao, Faliang Chang, and Sanyuan Hu. An Anchor-Free Convolutional Neural Network for Real-Time Surgical Tool Detection in Robot-Assisted Surgery. *IEEE Access*, 8:78193–78201, 2020.

[11] Zijian Zhao, Zhaorui Chen, Sandrine Voros, and Xiaolin Cheng. Real-time tracking of surgical instruments based on spatio-temporal context and deep learning. *Computer Assisted Surgery*, 24(sup1):20–29, 2019.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks For Biomedical Image Segmentation. *Lecture Notes in Computer Science*, 9351:234–241, 2015.

[13] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2015.

[14] Luis C. García-Peraza-Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. *arXiv*, pages 1–12, 2016.

[15] Luis C. Garcia-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sebastien Ourselin. ToolNet: Holistically-nested real-time segmentation of robotic surgical tools. *IEEE International Conference on Intelligent Robots and Systems*, 2017-Septe:5717–5722, 2017.

[16] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, 2017-Janua:3373–3378, 2017.

[17] S. M.Kamrul Hasan and Cristian A. Linte. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instrument. *arXiv*, pages 7205–7211, 2019.

[18] Alexey A. Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pages 624–628, 2019.

[19] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep Residual Learning for Instrument Segmentation in Robotic Surgery. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11861 LNCS:566–573, 2019.

[20] Emanuele Colleoni, Sara Moccia, Xiaofei Du, Elena De Momi, and Danail Stoyanov. Deep Learning Based Robotic Tool Detection and Articulation Estimation with Spatio-Temporal Layers. *IEEE Robotics and Automation Letters*, 4(3):2714–2721, 2019.

[21] Zhigang Li. CDPN : Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. pages 7678–7687.

[22] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2938–2946, 2015.

[23] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *arXiv*, 2017.

[24] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6D pose estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:7667–7676, 2019.

[25] Yannick Bukschat and Marcus Vetter. EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. *arXiv*, 2020.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:21–37, 2016.

[27] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:1530–1538, 2017.

[28] Iro Laina, Nicola Rieke, Christian Rupprecht, Josué Page Vizcaíno, Abouzar Eslami, Federico Tombari, and Nassir Navab. Concurrent Segmentation And Localization For Tracking Of Surgical Instruments. *Lecture Notes in Computer Science*, 10434 LNCS:664–672, 2017.

[29] Thomas Kurmann, Pablo Marquez Neila, Xiaofei Du, Pascal Fua, Danail Stoyanov, Sebastian Wolf, and Raphael Sznitman. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. *arXiv*, 1:505–513, 2017.

[30] Xiaofei Du, Thomas Kurmann, Ping Lin Chang, Maximilian Allan, Sebastien Ourselin, Raphael Sznitman, John D. Kelly, and Danail Stoyanov. Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE Transactions on Medical Imaging*, 37(5):1276–1287, 2018.

[31] Mert Kayhan, Okan Köpüklü, Mhd Hasan Sarhan, Mehmet Yigitsoy, Abouzar Eslami, and Gerhard Rigoll. Deep Attention Based Semi-Supervised 2D-Pose Estimation for Surgical Instruments. *ArXiv*, 2019.

[32] Masakazu Yoshimura, Murilo M. Marinho, Kanako Harada, and Mamoru Mitsuishi. Single-shot pose estimation of surgical robot instruments' shafts from monocular endoscopic images. *arXiv*, pages 9960–9966, 2020.

[33] Masakazu Yoshimura, Murilo Marques Marinho, Kanako Harada, and Mamoru Mitsuishi. MBAPose: Mask and Bounding-Box Aware Pose Estimation of Surgical Instruments with Photorealistic Domain Randomization. *ArXiv*, 2021.

[34] Md Kamrul Hasan, Lilian Calvet, Navid Rabbani, and Adrien Bartoli. Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Medical Image Analysis*, 70:101994, 2021.

[35] Spyridon Souipas, Anh Nguyen, Stephen Laws, Brian Davies, and Ferdinando Rodriguez Y Baena. 3dStool - A 3D Surgical Tool dataset for detection and pose estimation. https://github.com/SpyrosSou/3dStool, 2 2023.

[36] Aleksandr V Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-ICP. *Robotics: Science and Systems*, 2(4):435, 2009.

[37] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2017.

[38] Gioele Ciaparrone, Francesco Bardozzo, Mattia Delli Priscoli, Juanita Londono Kallewaard, Maycol Ruiz Zuluaga, and Roberto Tagliaferri. A comparative analysis of multi-backbone Mask R-CNN for surgical tools detection. *Proceedings of the International Joint Conference on Neural Networks*, 2020.

[39] Bo Chen, Tat Jun Chin, and Marius Klimavicius. Occlusion-Robust Object Pose Estimation with Holistic Representation. In *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 2223–2233. Institute of Electrical and Electronics Engineers Inc., 2022.

[40] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021.

[41] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, and Kalle Åstr. Revisiting the PnP Problem : A Fast , General and Optimal Solution. *IEEE International Conference on Computer Vision*, 2013.