In gls{rmis}, a marker-based method involves placing artificial markers on surgical instruments to aid vision-based instrument tracking\cite{villani2021development}. These markers can often be easily recognised in complex surgical environments. However, if the markers are obscured, damaged, or obscured by blood coverage, this may result in detection failure\cite{ma2021comprehensive}. In addition, markers on the surface of surgical instruments must meet sterility requirement\cite{xu2023graph}. To address these issues, marker-less methods have been gradually proposed\cite{reilink20133d}. marker-less methods do not rely on artificial markers in endoscopic procedures, but rather on natural features of the surgical instruments for gesture estimation. This method does not require additional marking process for the instruments and is able to adapt to various environmental changes with higher flexibility. However, the current marker-less method still faces some challenges, such as being susceptible to interference from lighting conditions, blood occlusion, and instrument reflections, and may not perform as consistently as the marker-based method in complex scenes\cite{hein2021towards}.

## 2. Related Works

## \subsection{Traditional non-learning methods}

Traditional non-learning pose estimation methods are based on geometric modelling, algebraic techniques, and computer vision approaches\cite{fan2024reinforcement}. Non-learning methods differ from deep learning methods, which require a large amount of data as input. Typically, the core of traditional non-deep learning methods is featuring extraction, which is used to obtain a unique representation of an object by identifying edges, keypoints, or regional features\cite{fan2024reinforcement}. In the early days, the main methods for feature recognition include \gls{sift}\cite{lakshmi2017image}, which is a classical local feature descriptor that helps machines to identify and match feature points in different images, and to find key points in different scale spaces. Another is the \gld{surf}\cite{wijesinghe2010speed}, which is an improved variant of \gls{sift} that improves the performance of feature extraction by optimising the process of feature detection and description.

After feature extraction is completed, the target model needs to be parameterised. The \gls{pnp}\cite{[yun2017object,lu2018review} algorithm is a commonly used algorithm for model parameterisation, which uses a set of 3D points in the scene and their 2D projections in the image to estimate translation and rotation parameters. Another method is the \gls{icp} algorithm\cite{bellekens2014survey}, which computes the pose relationship between two-point clouds by minimising the distance between corresponding points.

Non-learning methods have better interpretability and can achieve more accurate results while saving computational resources. However, such methods are less robust when dealing with complex scenes and lighting changes and have certain limitations\cite{bellekens2014survey}.

## \subsection{Deep Leaning Method}

In recent years, with the development of deep learning, it has been gradually applied to the field of surgical instrument pose estimation. Different from the traditional way that relies on geometric models and manual feature extraction, deep learning methods can infer the complex relationship between points and points from a large amount of data \cite{ bellekens2014survey}, and according to the facts, deep learning methods are more capable of handling complex scenes with lighting changes\cite{fan2024reinforcement}. Through the concern classification of the model, it can be divided into Holistic method and Intermediate representation method.

### \subsection{2.1.1 Holistic}

The Holistic method extracts estimated surgical instrument poses by modelling global features of the entire scene\cite{fan2024reinforcement}. This approach does not rely on local detail features, but rather extracts pose information from global features making the Holistic method highly robust to complex scene variations. In 2015, Alex Kendall and his team proposed PoseNet, a deep learning method that directly regresses camera pose from monocular RGB images, enabling end-to-end position and orientation end-to-end estimation\cite{kendall2015posenet}. In 2020, Yannick Bukschat et al. proposed EfficientPose, an end-to-end 6D multi-target pose estimation method. The model is capable of simultaneously detecting the 2D bounding boxes of multiple targets in a monocular image and regressing their complete 6D poses in 3D space\cite{bukschat2020efficientpose}. In 2022, Bo Chen and colleagues developed the ROPE framework, which introduces a new occlusion enhancement technique and a multi-precision supervised mechanism, aiming to learn deep features that are robust to occluded environments, thus improving the accuracy of pose estimation in object-occluded scenes\cite{chen2022occlusion}.

The Holistic method is able to capture the overall features of an object directly from the whole image, with a low dependence on feature points, without the need for precise positioning of feature points or additional feature extraction steps, which makes the model structure more concise\cite{chen2022occlusion}. However, the Holistic method is less accurate in dealing with local details, and when surgical instruments are occluded, it is difficult to recover the occluded instrument information from the overall features\cite{watson2014nature}.

### \subsection{2.1.2 Intermediate Representations}

The Intermediate Representation method decomposes the complex pose estimation task into multiple more manageable subtasks by introducing a finer-grained intermediate description of the target. By extracting local features, the method effectively solves the problem that it is difficult to accurately estimate the pose of surgical instruments when they are occluded\cite{song2020hybridpose}.

In 2017, Yu Xiang and his team proposed PoseCNN for pose estimation in complex scenes. The method decomposes the pose estimation task into multiple components that deal with 3D translations and rotations of images separately. In addition, PoseCNN introduces a novel loss function that allows the network to better handle objects with symmetry\cite{xiang2017posecnn}.

Subsequently, in 2019, Sida Peng and his team proposed PVNet\cite{peng2019pvnet}. this approach uses a pixel-level voting network that significantly improves pose estimation accuracy in occluded and truncated scenes by predicting vectors from each pixel to a key point, combined with a RANSAC-based voting mechanism.

In 2020, Masakazu Yoshimura and his team developed a deep learning model based on an improved SSD-6D architecture\cite{yoshimura2020single}. The model utilises a manually generated dataset of single-frame endoscopic images combined with data enhancement techniques to effectively address occlusion and perspective distortion problems common in surgical environments.

In 2022, Mitchell Doughty and his team proposed HMD-EgoPose\cite{yoshimura2020single}. The method uses the EfficientDet-D0 network for multi-scale feature extraction and combines rotational, translational, and hand sub-networks to achieve 6-degree-of-freedom markerless pose estimation in monocular RGB images.

In 2024, Jihun Park and his team introduced a new occlusion-aware loss function based on the YOLOv8 model, which dramatically improved the accuracy of precise detection and pose estimation of key points of surgical instruments in complex occlusion environments\cite{park2024towards}. The research team trained the model on a real surgical dataset, which significantly improved its robustness in real surgical scenarios.

The Intermediate representation method makes the task much less difficult by decomposing the complex pose estimation task into multiple, more manageable subtasks. At the same time, Intermediate Representation models local features so that the model can still maintain high stability in complex scenes. However, this method requires high accuracy in data labelling, and the accumulation of errors may affect the accuracy of the final results due to the inclusion of multiple intermediate steps\cite{xu2023graph,allan20183}.

# Reference

@inproceedings{villani2021development,

title={Development of an Augmented Reality system based on marker tracking for robotic assisted minimally invasive spine surgery},

author={Villani, Francesca Pia and Di Cosmo, Mariachiara and Simonetti, {\'A}lvaro Bertelsen and Frontoni, Emanuele and Moccia, Sara},

booktitle={International Conference on Pattern Recognition},

pages={461--475},

year={2021},

organization={Springer}

}

@article{ma2021comprehensive,

title={Comprehensive review of surgical microscopes: technology development and medical applications},

author={Ma, Ling and Fei, Baowei},

journal={Journal of biomedical optics},

volume={26},

number={1},

pages={010901--010901},

year={2021},

publisher={Society of Photo-Optical Instrumentation Engineers}

}

@inproceedings{xu2023graph,

title={Graph-based Pose Estimation of Texture-less Surgical Tools for Autonomous Robot Control},

author={Xu, Haozheng and Runciman, Mark and Cartucho, Jo{\~a}o and Xu, Chi and Giannarou, Stamatia},

booktitle={2023 IEEE International Conference on Robotics and Automation (ICRA)},

pages={2731--2737},

year={2023},

organization={IEEE}

}

@article{reilink20133d,

title={3D position estimation of flexible instruments: marker-less and marker-based methods},

author={Reilink, Rob and Stramigioli, Stefano and Misra, Sarthak},

journal={International journal of computer assisted radiology and surgery},

volume={8},

pages={407--417},

year={2013},

publisher={Springer}

}

@article{hein2021towards,

title={Towards markerless surgical tool and hand pose estimation},

author={Hein, Jonas and Seibold, Matthias and Bogo, Federica and Farshad, Mazda and Pollefeys, Marc and F{\"u}rnstahl, Philipp and Navab, Nassir},

journal={International journal of computer assisted radiology and surgery},

volume={16},

pages={799--808},

year={2021},

publisher={Springer}

}

@article{fan2024reinforcement,

title={A Reinforcement Learning Approach for Real-Time Articulated Surgical Instrument 3D Pose Reconstruction},

author={Fan, Ke and Chen, Ziyang and Liu, Qiaoling and Ferrigno, Giancarlo and De Momi, Elena},

journal={IEEE Transactions on Medical Robotics and Bionics},

year={2024},

publisher={IEEE}

}

@article{fan2024reinforcement,

title={A Reinforcement Learning Approach for Real-Time Articulated Surgical Instrument 3D Pose Reconstruction},

author={Fan, Ke and Chen, Ziyang and Liu, Qiaoling and Ferrigno, Giancarlo and De Momi, Elena},

journal={IEEE Transactions on Medical Robotics and Bionics},

year={2024},

publisher={IEEE}

}

@article{lakshmi2017image,

title={Image registration techniques based on the scale invariant feature transform},

author={Lakshmi, K Divya and Vaithiyanathan, V},

journal={IETE Technical Review},

volume={34},

number={1},

pages={22--29},

year={2017},

publisher={Taylor \& Francis}

}

@article{wijesinghe2010speed,

title={Speed up Robust Features in Computer Vision Systems},

author={Wijesinghe, WOKAS},

year={2010}

}

@inproceedings{yun2017object,

title={Object recognition and pose estimation for modular manipulation system: Overview and initial results},

author={Yun, Woo-han and Lee, Jaeyeon and Lee, Joo-Haeng and Kim, Jaehong},

booktitle={2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)},

pages={198--201},

year={2017},

organization={IEEE}

}

@inproceedings{lu2018review,

title={A review of solutions for perspective-n-point problem in camera pose estimation},

author={Lu, Xiao Xin},

booktitle={Journal of Physics: Conference Series},

volume={1087},

number={5},

pages={052009},

year={2018},

organization={IOP Publishing}

}

@inproceedings{bellekens2014survey,

title={A survey of rigid 3d pointcloud registration algorithms},

author={Bellekens, Ben and Spruyt, Vincent and Berkvens, Rafael and Weyn, Maarten},

booktitle={AMBIENT 2014: the Fourth International Conference on Ambient Computing, Applications, Services and Technologies, August 24-28, 2014, Rome, Italy},

pages={8--13},

year={2014}

}

@article{lecun2015deep,

title={Deep learning},

author={LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey},

journal={nature},

volume={521},

number={7553},

pages={436--444},

year={2015},

publisher={Nature Publishing Group UK London}

}

@inproceedings{valderrama2022towards,

title={Towards holistic surgical scene understanding},

author={Valderrama, Natalia and Ruiz Puentes, Paola and Hern{\'a}ndez, Isabela and Ayobi, Nicol{\'a}s and Verlyck, Mathilde and Santander, Jessica and Caicedo, Juan and Fern{\'a}ndez, Nicol{\'a}s and Arbel{\'a}ez, Pablo},

booktitle={International conference on medical image computing and computer-assisted intervention},

pages={442--452},

year={2022},

organization={Springer}

}

@inproceedings{kendall2015posenet,

title={Posenet: A convolutional network for real-time 6-dof camera relocalization},

author={Kendall, Alex and Grimes, Matthew and Cipolla, Roberto},

booktitle={Proceedings of the IEEE international conference on computer vision},

pages={2938--2946},

year={2015}

}

@article{bukschat2020efficientpose,

title={EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach},

author={Bukschat, Yannick and Vetter, Marcus},

journal={arXiv preprint arXiv:2011.04307},

year={2020}

}

@inproceedings{chen2022occlusion,

title={Occlusion-robust object pose estimation with holistic representation},

author={Chen, Bo and Chin, Tat-Jun and Klimavicius, Marius},

booktitle={Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision},

pages={2929--2939},

year={2022}

}

@inproceedings{chen2022occlusion,

title={Occlusion-robust object pose estimation with holistic representation},

author={Chen, Bo and Chin, Tat-Jun and Klimavicius, Marius},

booktitle={Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision},

pages={2929--2939},

year={2022}

}

@misc{watson2014nature,

title={The nature of holistic processing in face and object recognition: Current opinions},

author={Watson, Tamara L and Robbins, Rachel A},

journal={Frontiers in Psychology},

volume={5},

pages={3},

year={2014},

publisher={Frontiers Media SA}

}

@inproceedings{song2020hybridpose,

title={Hybridpose: 6d object pose estimation under hybrid representations},

author={Song, Chen and Song, Jiaru and Huang, Qixing},

booktitle={Proceedings of the IEEE/CVF conference on computer vision and pattern recognition},

pages={431--440},

year={2020}

}

@article{xiang2017posecnn,

title={Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes},

author={Xiang, Yu and Schmidt, Tanner and Narayanan, Venkatraman and Fox, Dieter},

journal={arXiv preprint arXiv:1711.00199},

year={2017}

}

@inproceedings{peng2019pvnet,

title={Pvnet: Pixel-wise voting network for 6dof pose estimation},

author={Peng, Sida and Liu, Yuan and Huang, Qixing and Zhou, Xiaowei and Bao, Hujun},

booktitle={Proceedings of the IEEE/CVF conference on computer vision and pattern recognition},

pages={4561--4570},

year={2019}

}

@inproceedings{yoshimura2020single,

title={Single-shot pose estimation of surgical robot instruments’ shafts from monocular endoscopic images},

author={Yoshimura, Masakazu and Marinho, Murilo M and Harada, Kanako and Mitsuishi, Mamoru},

booktitle={2020 IEEE International Conference on Robotics and Automation (ICRA)},

pages={9960--9966},

year={2020},

organization={IEEE}

}

@article{doughty2022hmd,

title={HMD-EgoPose: Head-mounted display-based egocentric marker-less tool and hand pose estimation for augmented surgical guidance},

author={Doughty, Mitchell and Ghugre, Nilesh R},

journal={International journal of computer assisted radiology and surgery},

volume={17},

number={12},

pages={2253--2262},

year={2022},

publisher={Springer}

}

@inproceedings{park2024towards,

title={Towards Precise Pose Estimation in Robotic Surgery: Introducing Occlusion-Aware Loss},

author={Park, Jihun and Hong, Jiuk and Yoon, Jihun and Park, Bokyung and Choi, Min-Kook and Jung, Heechul},

booktitle={International Conference on Medical Image Computing and Computer-Assisted Intervention},

pages={639--648},

year={2024},

organization={Springer}

}

@article{allan20183,

title={3-D pose estimation of articulated instruments in robotic minimally invasive surgery},

author={Allan, Max and Ourselin, S{\'e}bastien and Hawkes, David J and Kelly, John D and Stoyanov, Danail},

journal={IEEE transactions on medical imaging},

volume={37},

number={5},

pages={1204--1213},

year={2018},

publisher={IEEE}

}

# Glossary：

\newacronym{sift}{SIFT}{Scale Invariant Feature Transform}

\newacronym{surf}{SURF}{Speeded Up Robust Feature}

\newacronym{icp}{ICP}{Iterative Nearest Point}