Globally Learnable Point Set Registration Between 3D CT and Multi-view 2D X-ray Images of Hip Phantom

Jin Pan[1], Zhe Min[2], Ang Zhang[1], Han Ma[1], and Max Q.-H. Meng[*1,3]

***Abstract*— 2D-3D registration is a crucial step in Image-Guided Intervention, such as spine surgery, total hip replacement, and kinematic analysis. To find the information in common between pre-operative 3D CT images and intra-operative X-ray 2D images is vital to plan and navigate. In a nutshell, the goal is to find the movement and rotation of the 3D body's volume to make them reorient with the patient body in the 2D image space. Due to the loss of dimensionality and different sources of images, efficient and fast registration is challenging. To this end, we propose a novel approach to incorporate a point set Neural Network to combine the information from different views, which enjoys the robustness of the traditional method and the geometrical information extraction ability. The pre-trained Deep BlindPnP captures the global information and local connectivity, and each implementation of view-independent Deep BlindPnP in different view pairs will select top-priority pairs candidates. The transformation of different viewpoints into the same coordinate will accumulate the correspondence. Finally, a POSEST-based module will output the final 6 DoF pose. Extensive experiments on a real-world clinical dataset show the effectiveness of the proposed framework compared to the single view. The accuracy and computation speed are improved by incorporating the point set neural network.**

I. INTRODUCTION

Different modalities, dimensions, and viewpoints of the same object of a scene are various in appearance, the intuition is they have some hidden correspondence. In this work, we study how we determine the relation among the same object's images, in 3D dimension and multi-view 2D dimension, namely multi-view 2D-3D registration.

The 2D-3D registration is widely applied in Autonomous Vehicles [1], [2], Image-Guided Intervention [3], and Robotics [4]. In Image-guided Intervention of Hip Joint [5], pre-operative 3D CT is implemented to provide the patient's information to plan the surgery [6], [7]. While in the intra-operative stage, the 2D X-rays from different viewpoints are used to observe the updated state of the patient, for the navigation of the surgical instrument.

[1]Jin Pan, Ang Zhang, Han Ma are with the Robotics, Perception and AI Lab, Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong SAR, China.

[2]Zhe Min is with Department of Medical Physics and Biomedical Engineering, and also with Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, United Kingdom.

[3]Max Q.-H. Meng is with the Department of Electronic and Electrical Engineering of the Southern University of Science and Technology in Shenzhen, China, on leave from the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the Shenzhen Research Institute of the Chinese University of Hong Kong in Shenzhen, China.

[*] Max Q.-H. Meng is the corresponding author.

To fuse the information is a useful tool for visualization and monitoring. The input of the problem is 3d and 2d image modalities, with underlying connection, i.e., capturing the same patient from different types of medical equipment. The objective is to search for the rigid motion for a good mapping between the aligned images so that 3D images' projection best aligns with the 2D images. The different modalities and dimensions of CT and X-ray, making the registration problem challenging.

With ground truth, the 2D images and 3D images will be well organized in order. However, we face a problem of the unordered situation with infinite solutions. To convert the disorder into order, there are many previous attempts in the community. The related literature can be divided into two branches: Optimization-based methods [8]–[10] and learning-based methods [11]–[14]. Optimization-based methods usually model the 2d-3d registration problem into how to minimize an objective error function which describes the mismatch between the 2 point sets. In another direction, learning-based methods usually solve the problem in an end-to-end manner. They build a learning-based method to capture different-level features, learning the parameters with the training data.

The optimization-based methods are easier to understand and the result can be guaranteed, while the computation is heavy, therefore, is hard to be implemented in real-world scenarios. The learning-based methods usually take up less time for testing, while the lack of training data and robustness limits these methods. As a result, we make use of the speed of the learning-based method to speed up the optimization-based method, meanwhile making the result reasonable. The intuition is to incorporate a learning-based method into a traditional optimization framework [15], guaranteeing the robustness and computation time.

In this work, we explore how to incorporate a novel learning-based method Deep BlindPnP [16], [17] into a traditional framework, and combine all the information from different views. The entire framework of the proposed method is as follows: Each point set will undertake feature extraction by a Pointnet-based module. The distance in feature space between 3D point set and each 2D point set is calculated, pair by pair. After calculation, the top possible pair candidate will be accumulated, to vote for a final result. The outputting 6 DoF pose by adapted POSEST [18] will make the 3D volume's projection matchable with 2D point set.

We evaluate our method in the public Hip Joint Dataset [5], the additional views improve the accuracy compared to the single-view setting.

The main contribution of our paper is two-fold:

- We explore the Globally learnable 2D-3D Point Set Registration in multi-view settings.
- We implement the method in the real-world clinical dataset, hip joint dataset. The images captured from different views can speed up the convergence of searching and improve the accuracy.

II. Related Work

To fuse images from different modalities, the key salient points such as boundary and edge are selected to represent the image. Point set registrationčitepomerleau2015review is utilized to merge multiple data source or map new measurement to prior model [19].

From the perspective of dimension, the point set registration can be generally divided into 3D-3D point set registration [20], [21], 3D-2D registration [3].

The scenarios of this work focus on the 2D-3D registration. In general, the 2D-3D registration methods consist of optimization-based methods and learning-based methods.

For optimization category [22], the direction is to define an objective function on how the 3D images' projection aligns with 2D images, then optimize it. As the setting of single view [23] is ill-posed, other methods extend the method into multi-view setting [24], [25], improving the accuracy and speed.

For learning-based algorithms [11], several novel deep learning approaches are proposed to match the 3D-2D images. The 2d3d-matchnet [4] is proposed to build an end-to-end deep network to jointly learn the feature descriptors for keypoint of 2D image and 3D point cloud. Deep BlindPnP [16], [26] is proposed to end-to-end output pose given 2d-3d point sets. But these methods are hungry for training data, which is not applicable in medical scenarios. Here the pre-trained model is implemented to capture the local geometry information of point set.

Other than dimension, we will review the point set registration viewpoint setting. The typical problem is single-view [27] where the capturing of images is easy to realize. In the case of easy acquisition of multi-view setting or small capture angle [28], the problem can be extended. Multi-view Point-To-Plane correspondence model [29] extends the PPC model. POINT^2 [14] makes use of Neural Network to find 2D point-to-point correspondences by tracking a small number of 3D POI.

III. Problem Formulation And Methodology

This work focuses on a global 2D-3D Point set rigid registration problem. Taking a 3D volume and several 2D images as input, acquiring the visual information of the same scene or object, a desired 6 DoF pose can be the bridge between the two modalities. With the transformation of 6 DoF pose, the projection of transformed 3D data can be aligned well by the 2d images, pair to pair.

Reviewing the problem, the potential candidates are inexhaustible, and the geometric property of rotation makes the computation heavy. The key challenge is to find the candidate globally in an efficient method.

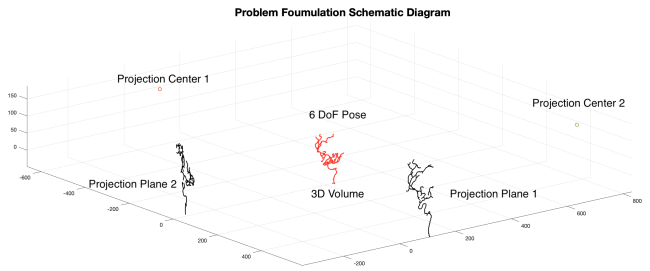

Fig. 1: The setting of registration between Multiple 2D images and 3D image. The optimal pose between the initial red 3D points with the projected 2D points will make the overlay of 3D points(in red) matching the 2D points (in black). The multi-view setting is simplified to only 2 views for visualization.

To make the challenging problem easier to handle, we treat the key points of the 3D volume and 2D images to represent the corresponding image modalities. The key geometric information, local and global, is reserved in the extracted point set.

We will formulate the problem in two steps, the single view firstly and then extend it into a multiple view setting.

A. 2D-3D Rigid Point Set Registration

In a camera coordinate system P, the 3D point set is $\mathcal{M} = \{m_j\}, j = 1, ..., M$, where $m_j \in \mathbb{R}^3$ are 3D point coordinates. And 2D point set $\mathcal{D} = \{d_i\}, i = 1, ..., N$ where $d_i \in \mathbb{R}^2$ are 2D point coordinates. The 3D point set and 2D point set are paired by an implicit 6D pose, a 3D rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$. In the ideal case, the projections of the transformed 3D points into different projection plane and 2D points should satisfy

$$P(R * M + t) \rightarrow D \tag{1}$$

To make the computation easier, we refer to Liu et al. [23] to calculate the cardinality of matching set:

$$Q(R, t) = \sum_i \max_j \mathbf{1}(||P(R * m_j + t) - d_i|| \leq \delta) \tag{2}$$

where $\mathbf{1}(\cdot)$ is an indicator function, the indicator function outputs 1 if the inside part is true, otherwise 0.

Therefore, the problem is transformed to search for the pose to maximize the cardinality of inlier set. The pair between the 3d point set and its inlier 2d points will be the correspondence.

The above definition of the problem is ideal, where the 3d point has a one-by-one paired 2D point. However, the practical scenarios are that the 3d points' projection is partially overlapped by the 2D points, or that the key points extracted in different modalities have no direct pairs. Therefore, there are no explicit and specific correspondences. During finding the optimal pose, the correspondence will be established.

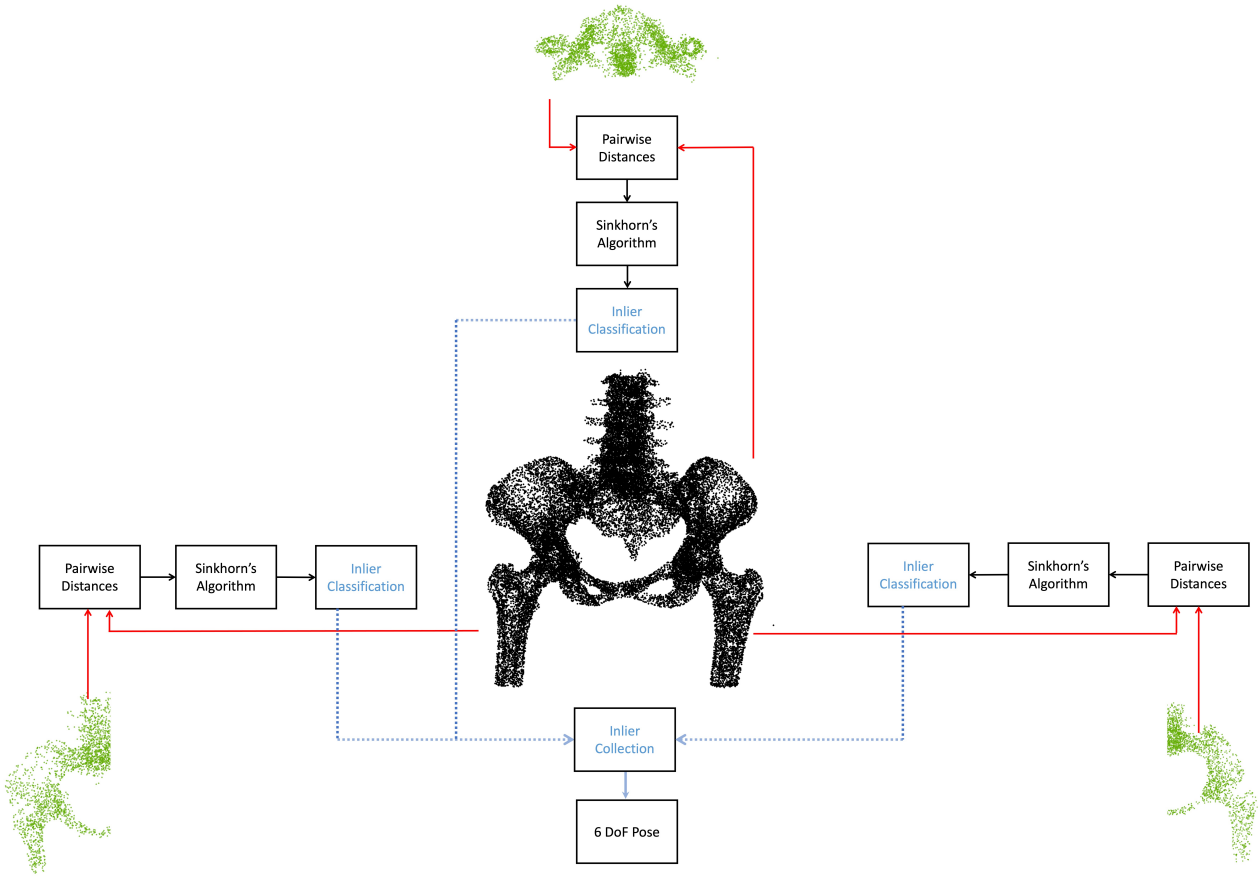

Fig. 2: The entire framework of the proposed method. Each point set will undertake feature extraction by a Pointnet-based module. The distance in feature space between the 3D point set and each 2D point set is calculated, pair by pair. After calculation, the top possible pair candidate will be selected to accumulate, to vote for a final result. The outputting 6 DoF pose by adapted POSEST [18] will make the 3D volume's projection matchable with 2D point set.

Here we define the correspondence of 3d point m_j as $CP(m_j)$:

$$CP(m_j) = \{d_i, i = 1, ..., N || P(R * m_j + t) - d_i || \leq \delta\} \quad (3)$$

where the points d_i has a distance less than a threshold δ with transformed m_j's projection.

B. Introduce New Viewpoints

Now we turn to the scenario of multi-view shown in Fig. 1, we need to add a notation v corresponding to each view.

$$P^v(R * M + t) \rightarrow D^v \quad (4)$$

Therefore, the correspondences optimization of has transformed as follows:

$$\max_{R,t} \sum_j \mathbf{1}(CP^v(m_j) \neq \emptyset) \quad s.t. R \in SO(3), t \in \mathbb{R}^3 \quad (5)$$

where the $CP^v(m_j)$ can be multiple points or null.

Therefore, the problem is transformed to search for the pose to maximize the cardinality of inlier set, i.e. the quality of the registration.

So far the objective and standard to evaluate how the two point sets are connected, is formulated shown in Fig. 1. In the next section, we will introduce how to search for the optimal 6 DoF pose.

C. Methodology

With the objective defined in the last section, we need to find the optimal pose to optimize the total number of matchable sets' cardinality.

As depicted in Fig. 2, the entire framework of the proposed method is as follows: Each point set will undertake feature extraction by Pointnet-based module. The distance in feature space between 3D point set and each 2D point set is calculated, pair by pair. After calculation, the top possible pair candidate will be accumulated, to vote for a final result. The outputting 6 DoF pose by adapted POSEST [18] will make the 3D volume's projection matchable with 2D point set.

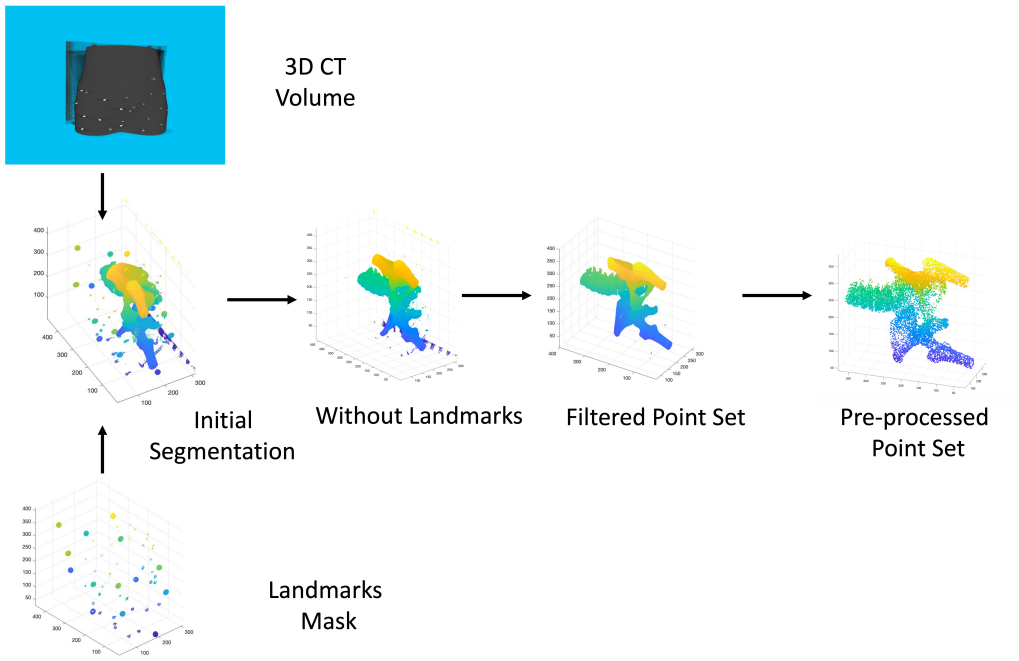

Fig. 3: The extraction of 3D point set from 3D CT volume. The 3D volume is initially segmented to output a surface of hip joint, then the mask of fiducial landmarks (for gold standard) is used to prune the landmark volumes. The noise is filtered and then downsampled, outputting the preprocessed point cloud.

The Deep BlindPnP neural network [16], [17] is a neural network to output the 6 DoF pose of the 3D point set to make its projection align with the 2D point set. While it is not applicable in the medical scenario for lack of data. In this work, we want to use the pre-trained model to accelerate the registration of different data distributions.

1) Deep BlindPnP Algorithm: In this section, we will introduce Deep BlindPnP, the important part to speed up backbone BnB. This is the first neural network to end-to-end output the 6 DoF pose given 2D points and 3D points.

The intuition of Deep BlindPnP is that the 2D and 3D structures share similar features although in different dimensions. Therefore, the input 2D and 3D point sets are fed into PointNet-based feature extractor, and the correspondence matrix is calculated by Deep Declarative Networks [30]. The top possible correspondence (inline) will be treated as known and correct correspondence, now the problem is collapsed into a classic PnP problem.

The input is 2D and 3D point set, the output will be end-to-end 6 DoF pose. It is worth mentioning that the optimal pose acquired by a pre-trained neural network, is likely to be optimal pose without any guarantee.

2) Extension to Multiple View Setting: With different viewpoints, each view-independent setting will undertake the same feature extraction and inlier classification by pre-trained Deep BlindPnP neural network, then the candidate pseudo known pairs of different views are transformed into the same coordinate system.

The entire framework of the proposed method is as follows: Each point set will undertake feature extraction by Pointnet-based module. The distance in feature space between 3D point set and each 2D point set is calculated, pair by pair. After calculation, the top possible pair candidate will be accumulated, to vote for a final result. The outputting 6 DoF pose by adapted POSEST [18] will make the 3D volume's projection matchable with 2D point set.

The pre-trained Deep BlindPnP neural network of corresponding 2D-3D set pairs will output the potential 2D-3D point pairs. With the candidate pairs in different views outputting the previous step, the candidate pair will be transformed into the same projection setting depicted in Fig. 1. Finally, a combination of different point pairs will output a 6 DoF pose by POSEST.

IV. Experiments

In this section, we show the performance of experiments in a public clinical dataset, Hip Joint gold-standard Dataset [5]. This solid work collects a full scan computed tomography of a female patient. The capture site is a hip phantom.

And 19 2D X-Ray Images are collected in different views. And fiducial marks are used to offer the gold standard. We conduct the experiment to prove that the matchable error and computation time of our proposed method can be improved compared with the single settings.

We pre-process the 3D images and 2D images, segment and extract the surface points, which model the shape information of volume and images. We feed it into the pre-trained end-to-end neural network, then implement the result from different views' 2D-3D sets pairs to speed up the computation.

A. Pre-processing

The original data consists of one CT and 19 viewpoint groups and different views of a phantom with landmarks. The data is noisy, which has to undertake preprocessing. As illustrated in Fig. 3, the 3D images are segmented, then extracted as a surface point set. Now that the bone is rigid, there is no motion between the bone and the surface will represent the image.

Firstly, initial segmentation is conducted, then the mask of landmarks is used to prune the noise of markers. Finally, we downsample to acquire a more sparse point set. It is worth mentioning that we don't need the input point to be sparse in our scenario. The training of Deep BlindPnP takes up a long time, while testing needs only 1-2s.

The number of the extracted 3D points is 1372, and the number of the extracted 2D points of four views ranged from 411 to 606.

B. Implementation Details

In this section, we will introduce how we select the hyper-parameters empirically. After three times' test, the average inline number and computation time of different methods is evaluated.

The Deep BlindPnP is designed for end-to-end 2d-3d point set registration, and the feature extraction part is based on PointNet. Therefore, it needs enough training data to feed. In the medical scenario, the lack of medical data is very common. So we choose to feed our medical data into the pre-trained models. By empirically testing, we find the pre-trained data in ModelNet40 [31] and NYU-RGBD [32] work well. And the results of twice utilization, of Deep BlindPnP of these two models pre-trained on two datasets, are averaged.

And we implement the POSEST, we set PROBABILITY_CLOSE_TO_ONE as 0.991, MINIMUM_INLIERS_ADAPT as 0.07, MINIMUM_ITERATIONS_FRAC as 0.10, MIN_TRIANG _AREA as 150.0 , and for the RANSAC parameters RANSAC_OUTL _THRESH as 3.0.

C. Results

We compare the inline number and computation time of our methods and baseline methods. The quantitative results are depicted in Table I.

The number of 3D point set is 1372 in all cases, and the average 2D point set is 497.3, 510.2, 547.2, and 519.0 for four cases. The first experiment is ICP [33], a local method. It is very efficient, while very easy to fall in the local minimum. The second pre-trained BlindPnP is trained in other datasets. The computation time is 1.2s, while the accuracy is better. The percentage of inline is 0.20 and 0.50.

TABLE I: The average inline numbers and computation time performance.

	Total 3d, 2d numbers	Inline number	Time (s)
ICP	1372, 497.3	101.1 $\pm$ 10.8	0.4 $\pm$ 0.1
Pretrained BlindPnP	1372, 510.2	255.5 $\pm$ 50.2	1.2 $\pm$ 0.2
Two Views	1372, 547.2	370.1 $\pm$ 19.1	5.9 $\pm$ 0.4
Three Views	1372, 519.0	393.5 $\pm$ 13.8	5.7 $\pm$ 0.4

The above two methods take the single view as input. The Two View and Three View have a total of 547.2 and 519.0. The computation of these two methods is more than the formal two. And the percentage of inline is 0.68 and 0.76. The three views case outperforms in four experiments. And the computation time of ICP is the least, while the Three View can achieve the best inline percentage.

The experiment shows the effectiveness of our method. With the combined information provided by Deep BlindPnP, the POSEST can be sped up. The performance of inline numbers is improved.

V. CONCLUSIONS

In this work, we focus on how to register 3D point set of hip CT and several 2D X-rays shot in different viewpoints. A novel approach is proposed to incorporate a point set neural network to combine the information from different views, which enjoys the robustness of the traditional method and the geometrical information extraction ability. The pre-trained Deep BlindPnP captures the global information and local connectivity, and each implementation of view-independent Deep BlindPnP in different view pairs can accumulate the information. By transforming the different viewpoints' pairs into the same coordinate, the collection's pairs can be combined. Finally, a POSEST-based module will output the final 6 DoF pose. Extensive experiments on a real-world clinical dataset show the efficacy of the proposed framework compared to the single view. The accuracy and computation speed are improved by incorporating the point set neural network. In the future, we will go deep into different strategies to combine the view, e.g., alternating or selection of viewpoints to improve the performance.

ACKNOWLEDGEMENT

This project is supported by National Key R&D program of China with Grant No. 2019YFB1312400, and Hong Kong RGC TRS grant T42-409/18-R, Hong Kong RGC GRF grant # 14211420 and Hong Kong Health and Medical Research Fund (HMRF) under Grant 06171066 awarded to Prof. Max Q.-H. Meng.

REFERENCES

[1] S. A. Parkison, J. M. Walls, R. W. Wolcott, M. Saad, and R. M. Eustice, "2d to 3d line-based registration with unknown associations via mixed-integer programming," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 11 046–11 052.

[2] Y. Liu, G. Chen, and A. Knoll, "Globally optimal camera orientation estimation from line correspondences by bnb algorithm," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 215–222, 2020.

[3] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3d/2d registration methods for image-guided interventions," *Medical image analysis*, vol. 16, no. 3, pp. 642–661, 2012.

[4] M. Feng, S. Hu, M. H. Ang, and G. H. Lee, "2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4790–4796.

[5] F. DIsidoro, C. Chênes, S. J. Ferguson, and J. Schmid, "A new 2d-3d registration gold-standard dataset for the hip joint based on uncertainty modeling," *Medical Physics*, 2021.

[6] T. De Silva, A. Uneri, M. Ketcha, S. Reaungamornrat, G. Kleinszig, S. Vogt, N. Aygun, S. Lo, J. Wolinsky, and J. Siewerdsen, "3d–2d image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch," *Physics in Medicine & Biology*, vol. 61, no. 8, p. 3009, 2016.

[7] S. Yoon, C. H. Yoon, and D. Lee, "Topological recovery for non-rigid 2d/3d registration of coronary artery models," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105922, 2021.

[8] N. Baka, C. Metz, C. Schultz, L. Neefjes, R. J. van Geuns, B. P. Lelieveldt, W. J. Niessen, T. van Walsum, and M. de Bruijne, "Statistical coronary motion models for 2d+ t/3d registration of x-ray coronary angiography and cta," *Medical image analysis*, vol. 17, no. 6, pp. 698–709, 2013.

[9] J. Wang, R. Schaffert, A. Borsdorf, B. Heigl, X. Huang, J. Hornegger, and A. Maier, "Dynamic 2-d/3-d rigid registration framework using point-to-plane correspondence model," *IEEE transactions on medical imaging*, vol. 36, no. 9, pp. 1939–1954, 2017.

[10] R. Schaffert, J. Wang, P. Fischer, A. Maier, and A. Borsdorf, "Robust multi-view 2-d/3-d registration using point-to-plane correspondence model," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 161–174, 2019.

[11] S. Miao, Z. J. Wang, Y. Zheng, and R. Liao, "Real-time 2d/3d registration via cnn regression," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1430–1434.

[12] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[13] S. Miao, S. Piat, P. Fischer, A. Tuysuzoglu, P. Mewes, T. Mansi, and R. Liao, "Dilated fcn for multi-agent 2d/3d medical image registration," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] H. Liao, W.-A. Lin, J. Zhang, J. Zhang, J. Luo, and S. K. Zhou, "Multiview 2d/3d rigid registration via a point-of-interest network for tracking and triangulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 638–12 647.

[15] R. Schaffert, J. Wang, P. Fischer, A. Borsdorf, and A. Maier, "Learning an attention model for robust 2-d/3-d registration using point-to-plane correspondences," *IEEE transactions on medical imaging*, vol. 39, no. 10, pp. 3159–3174, 2020.

[16] L. Liu, D. Campbell, H. Li, D. Zhou, X. Song, and R. Yang, "Learning 2d-3d correspondences to solve the blind perspective-n-point problem," *arXiv preprint arXiv:2003.06752*, 2020.

[17] D. Campbell, L. Liu, and S. Gould, "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization," in *European Conference on Computer Vision*. Springer, 2020, pp. 244–261.

[18] M. Lourakis and X. Zabulis, "Model-based pose estimation for rigid objects," in *International conference on computer vision systems*. Springer, 2013, pp. 83–92.

[19] J. Pan, X. Mai, C. Wang, Z. Min, J. Wang, H. Cheng, T. Li, E. Lyu, L. Liu, and M. Q.-H. Meng, "A searching space constrained partial to full registration approach with applications in airport trolley deployment robot," *IEEE Sensors Journal*, 2020.

[20] G. K. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun, and P. L. Rosin, "Registration of 3d point clouds and meshes: A survey from rigid to nonrigid," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 7, pp. 1199–1217, 2012.

[21] Z. Min, J. Wang, J. Pan, and M. Q.-H. Meng, "Generalized 3-d point set registration with hybrid mixture models for computer-assisted orthopedic surgery: From isotropic to anisotropic positional error," *IEEE Transactions on Automation Science and Engineering*, 2020.

[22] D. Knaan and L. Joskowicz, "Effective intensity-based 2d/3d rigid registration between fluoroscopic x-ray and ct," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2003, pp. 351–358.

[23] Y. Liu, Y. Dong, Z. Song, and M. Wang, "2d-3d point set registration based on global rotation search," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2599–2613, 2018.

[24] J. Pan, Z. Min, A. Zhang, H. Ma, and M. Q.-H. Meng, "Multi-view global 2d-3d registration based on branch and bound algorithm," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 3082–3087.

[25] K. Fu, Y. Liu, and M. Wang, "Global registration of 3d cerebral vessels to its 2d projections by a new branch-and-bound algorithm," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 1, pp. 115–124, 2021.

[26] D. Campbell*, L. Liu*, and S. Gould, "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization," in *ECCV*, 2020, * equal contribution.

[27] D. Ruijters, B. M. ter Haar Romeny, and P. Suetens, "Vesselness-based 2d–3d registration of the coronary arteries," *International journal of computer assisted radiology and surgery*, vol. 4, no. 4, pp. 391–397, 2009.

[28] A. Uneri, Y. Otake, A. Wang, G. Kleinszig, S. Vogt, A. J. Khanna, and J. Siewerdsen, "3d–2d registration for surgical guidance: effect of projection view angles on registration accuracy," *Physics in Medicine & Biology*, vol. 59, no. 2, p. 271, 2013.

[29] R. Schaffert, J. Wang, P. Fischer, A. Maier, and A. Borsdorf, "Robust Multi-View 2-D/3-D Registration Using Point-To-Plane Correspondence Model," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 161–174, 2020.

[30] S. Gould, R. Hartley, and D. J. Campbell, "Deep declarative networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[31] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[32] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[33] K. S. Huang and M. M. Trivedi, "Robust real-time detection, tracking, and pose estimation of faces in video streams," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 965–968.