

Article

Weakly Supervised Pose Estimation of Surgical Instrument from a Single Endoscopic Image

Lihua Hu ¹, Shida Feng ^{1,2} and Bo Wang ^{2,*}

¹ College of Computer Sciences and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China; hlh@tyust.edu.cn (L.H.); s202120110672@stu.tyust.edu.cn (S.F.)

² State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: bo.wang@ia.ac.cn

Abstract: Instrument pose estimation is a key demand in computer-aided surgery, and its main challenges lie in two aspects: Firstly, the difficulty of obtaining stable corresponding image feature points due to the instruments' high refraction and complicated background, and secondly, the lack of labeled pose data. This study aims to tackle the pose estimation problem of surgical instruments in the current endoscope system using a single endoscopic image. More specifically, a weakly supervised method based on the instrument's image segmentation contour is proposed, with the effective assistance of synthesized endoscopic images. Our method consists of the following three modules: a segmentation module to automatically detect the instrument in the input image, followed by a point inference module to predict the image locations of the implicit feature points of the instrument, and a point back-propagatable Perspective-n-Point module to estimate the pose from the tentative 2D–3D corresponding points. To alleviate the over-reliance on point correspondence accuracy, the local errors of feature point matching and the global inconsistency of the corresponding contours are simultaneously minimized. Our proposed method is validated with both real and synthetic images in comparison with the current state-of-the-art methods.

Keywords: 6D pose estimation; weakly supervised learning; back-propagatable PnP



Citation: Hu, L.; Feng, S.; Wang, B. Weakly Supervised Pose Estimation of Surgical Instrument from a Single Endoscopic Image. *Sensors* **2024**, *24*, 3355. <https://doi.org/10.3390/s24113355>

Academic Editor: Iulian I. Iordachita

Received: 28 April 2024

Revised: 15 May 2024

Accepted: 21 May 2024

Published: 23 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Robot-assisted minimally invasive surgery has become an important field of rapid development due to its enormous potential for application, where surgical instrument pose estimation is a key demand. However, instrument pose estimation faces two significant challenges: difficulty in obtaining stable feature points for surgical instruments due to their lack of texture and high reflectivity of surface materials, and the absence of labeled pose data for surgical instruments in monocular endoscopic images. Due to the two difficulties mentioned above, current surgical instrument pose estimation techniques often rely on additional positioning markers or control points, or the use of binoculars and depth cameras to provide annotation information [1,2]. For example, in order to enhance the active research of surgical tool identification and tracking, Lajkó et al. [3] utilize the surgeon's visual feedback loop to compensate for robot surgery inaccuracies. Attanasio et al. [4] identify anatomical structures and provide contextual awareness in the surgical environment, while [5,6] improve robotic surgical precision with the help of recognition of surgical phases. Kitaguchi et al. [7] employed a convolutional neural network to achieve high-precision automatic recognition of surgical actions in laparoscopic images. Nguyen et al. [8] used a CNN-LSTM neural network model and an IMU sensor-based automatic evaluation system to classify and regress motion data in the JIGSAWS [9] dataset. Pan et al. [10] proposed an automatic surgical skill evaluation framework based on visual motion tracking and deep learning. That method introduced a Kernelized Correlation Filter (KCF) to capture key motion signals from the robot. In sum, although quite a number of

different approaches are reported in the field, such works usually involve unnecessary assistant devices to the surgery, which cause inconvenience, even obstacles, to the surgery, and also hamper the system's wide applicability to some extent.

Object pose estimation is a key issue in the field of computer vision. Deep learning-based object pose estimation methods are rapidly rising and have become the mainstream research direction because of their excellent adaptability. These methods are mainly divided into three categories: corresponding point-based methods, voting-based methods, and region-based methods. Here is a short review of such methods:

Corresponding point-based methods first define feature points on 3D models, then deduce the projection position of feature points [11,12] in the image through a regression network, and then calculate the posture by solving the Perspective-n-Point (PnP) problem for these feature points. Rad et al. [13] proposed the BB8 algorithm, which infers the image coordinates of eight bounding box corners of an object first and then estimates the pose of the object by solving the PnP algorithm. Similarly, the Yolo-6d method [14] borrowed the idea of You only look once (Yolo) [15] and used the Efficient Perspective-n-Point (EPnP) algorithm [16] to calculate the pose by predicting the projection of the eight corners of the target's 3D bounding box on the 2D image. CullNet [17] improved Yolo-6d, using multi-channel output and confidence prediction to enhance pose estimation robustness. Other works [18,19] were proposed with improved results based on a similar idea. Hu et al. [20] proposed a segmentation-driven pose estimation algorithm. This algorithm first generates a foreground segment mask image, splits the image into several small patches, estimates object pose by every image patch containing the foreground, and finally fuses all the estimation results into a single final result. Pix2Pose [21] uses an encoder-decoder and generative adversarial network to infer projection coordinates of dense 3D model points and then estimate object pose with these corresponding 2D-3D point pairs. Similarly, Dense Pose Object Detector (DPOD) [22] uses UVmap to represent the pixel-level correspondence between 3D points and 2D projections and improves pose estimation accuracy by introducing a post-optimization module. Coordinates-based Disentangled Pose Network (CDPN) [23] also estimates the object pose based on pixel-level 2D-3D correspondence by separating translation and rotation components. Lu et al. [24] proposed a hybrid pseudo-labeling method that performs self-training by leveraging globally consistent pose estimates. Wang et al. [25] introduced self-supervised training into the field of pose estimation using SE(3) Equivariance networks.

The voting-based pose estimation method combines the advantages of sparse and dense correspondences, utilizing a voting mechanism to determine the object's pose. Pixel-wise Voting Network (PVNet) [26] locates feature points through a voting mechanism, demonstrating strong occlusion robustness. Similarly, HybridPose [27] introduces new features into the voting method, improving pose estimation performance. He et al. [28] proposed the PVN-3D, leveraging the Hough voting scheme to robustly estimate object pose.

Region-based pose estimation methods, instead of using point matching pairs, utilize image region information to infer the object's pose. Do et al. [29] proposed a Deep-6DPose algorithm. This method first utilizes Mask Region-based Convolutional Neural Network (R-CNN) [30] to obtain the foreground segmentation region, and then adds a pose regression branch to regress the object's rotation matrix and translation vector. Liu et al. [31] proposed an indoor object pose estimation algorithm based on the modified Faster R-CNN [32] framework. Sundermeyer et al. [33] trained an auto-encoder to estimate the rotation matrix, first by taking the whole image as input, and then solved the translation vector separately. The Deep Iterative Matching (DeepIM) [34] algorithm uses a 3D model to render synthetic images at first, and then the original image, rendered image, and corresponding mask are inputted to the FlowNet network [35] to obtain decoupled translation and rotation vectors. Deng et al. [36] proposed a weakly supervised pose estimation method based on robot interaction. Wang et al. [37] proposed the Self6D algorithm, leveraging unlabeled Red Green Blue-Depth (RGB-D) data in a self-supervised manner by means of differentiable rendering.

Unlike other applications, the endoscopic image-based pose estimation of surgical instruments has special difficulties, as we mentioned at the very beginning of this work, which are, mainly, the lack of stable feature points of surgical instruments, the large reflection property, and the lack of labeled training data. To tackle such difficulties, we propose a weakly supervised neural network framework to estimate the pose from the surgical instrument contour in the endoscopic image and use the synthetical image contours and key feature points of the instrument to assist the network training. In particular, a 2D feature points inference module is proposed to predict the corresponding 2D feature points from the segmented instrument contour in the image, and a back-propagatable PnP network module [38] and a differentiable rendering module [39] are used to estimate the pose.

To train our network effectively, we employ a two-phase training strategy. We train the segmentation module, separated from other modules, with the help of TerausNet-16 [40]. Our focus is on training the other network modules after the segmentation network, which takes segmentation mask images as input and outputs pose estimation. In summary, the main contributions of this work are as follows: (1) We propose a weakly supervised pose estimation method for surgical instruments from single endoscopic images, assisted by synthesized endoscopic images of surgical instruments; (2) both local discrepancy errors of predicted feature points and the global inconsistency of the whole contour are taken into account in the training phase; (3) our proposed method outperforms many state-of-the-art methods with large real and synthetic data; and (4) we create a synthetic surgical instrument contour image dataset of about 50,000 samples with corresponding pose ground truth, providing data support for future research.

2. Materials and Methods

The method proposed in this paper is outlined in the following steps:

Firstly, we utilized an offline contour-based 3D reconstruction method to obtain 3D models of surgical instruments and selected 8 implicit feature points by the farthest point sampling method. Then, we imported these 3D models into Unreal Engine 5 (UE5) virtual engine to generate the pose labeled synthetic segmentation image training data.

Secondly, to effectively leverage the auto-generated training data, we utilized a surgical instrument image segmentation module to preprocess the real surgical scene image. This module employs a TerausNet-16 structure to infer the foreground mask image of surgical instruments from the real surgical scene images.

Finally, our instrument pose estimation network was trained by the auto-generated labeled data. The network consists of cascaded modules: a backbone module consisting of alternated cascaded channel-space tensor self-attention layers and multi-resolution convolution layers to infer the heat map style feature, an implicit points' projection coordinates regression module, and a back-propagated PnP (BPnP) and differentiable rendering pose estimation module.

The overall framework of the proposed method in this paper is shown in Figure 1.

In the following section, the main modules will be elaborated.

2.1. Synthetic Training Data Generation

To address the stable feature tracking challenge caused by the lack of texture and specular reflection of surgical instrument surfaces, our method uses the foreground mask image of an instrument to infer the instrument's pose. With the foreground mask image as an input of the pose estimation network, disturbances caused by the drastic change in image appearance can be eliminated, leading to more robust pose inference. We utilized UE 5 virtual engine to generate foreground mask image training data.

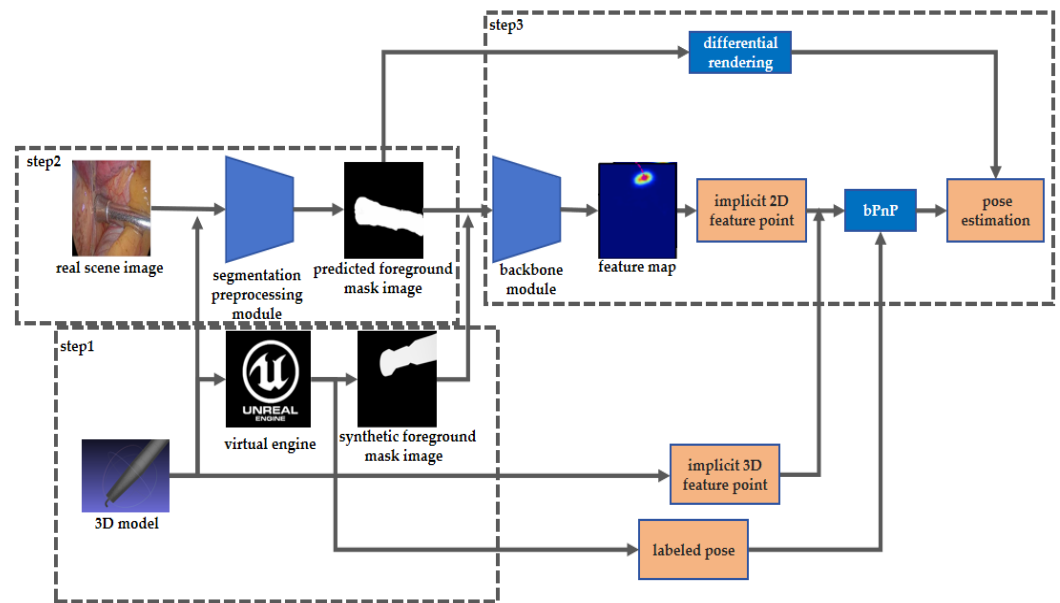


Figure 1. The overall framework of our weakly supervised surgical tools instrument pose estimation method. Three steps are involved. Step 1: Synthetic data generation. Step 2: Surgical instruments segmentation of real scenes. Step 3: Segmentation based surgical instruments pose estimation.

We used structure from the contour-based method to reconstruct the 3D models of three kinds of typical instruments: hooks, clippers, and suction, which are widely used in endoscopic minimally invasive surgery. After the 3D model reconstruction, eight implicit feature points were selected for each 3D model. Then, we imported the 3D models of instruments into UE 5 virtual engine, and we performed a uniform sampling within the reasonable range of the SE(3) pose space of the virtual camera. At each virtual camera pose sample point, a surgical instrument's foreground mask image was generated by virtual engine rendering, with the projections of implicit feature points being calculated and sampled pose parameters being recorded. We augmented the training dataset by adding noise to the generated images. The comparison between a captured real surgical instrument image and a rendered image with normal material and lighting in UE 5 virtual engine is shown in Figure 2. The samples of generated training image data are shown in Figure 3.

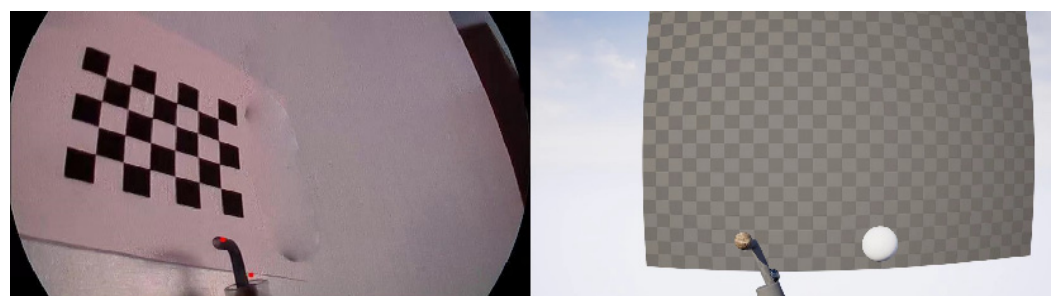


Figure 2. Comparison between a captured real surgical tool image and a rendered image with normal material and lighting in UE 5 virtual engine.

We generated the training dataset of three types of instruments: hooks, clippers, and suction. The dataset contains 50,000 of samples, and each data sample includes a synthetic foreground mask image, its corresponding pose label, the implicit feature points' 3D coordinates, and their 2D projections on the image. We utilized this dataset to train our model without further manual label effort, such that our model could be trained in a weakly supervised manner.

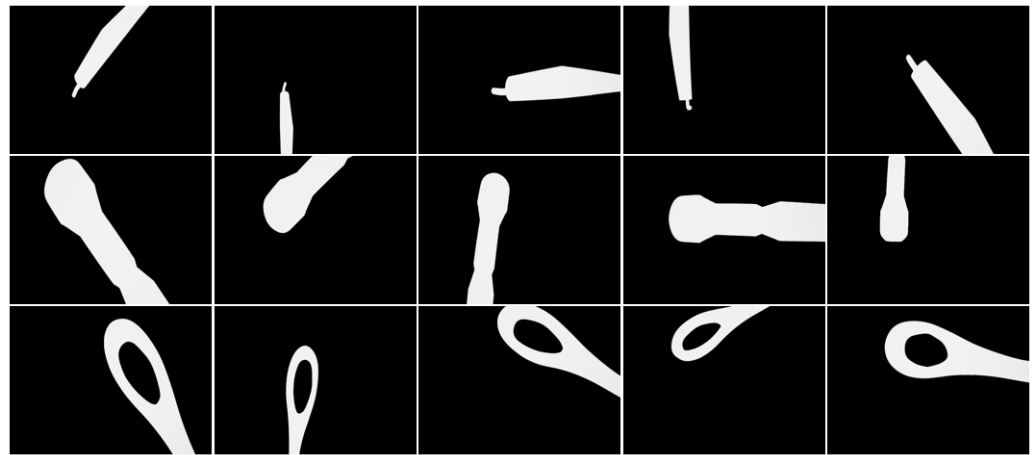


Figure 3. Auto-generated training dataset samples.

2.2. Instrument Segmentation Preprocessing

To better transfer the pose estimation network trained on synthetic data to estimate the pose of surgical instruments in real surgical scenes, we employed a surgical instrument segmentation network module before proceeding to our main pose estimation network. TernaNet-16 network, a popular one in the field of image segmentation, was used in this work. TernaNet-16 is an improvement and optimization based on the U-Net [41] architecture. Its model encoding layer replaces the original design with VGG16 [42], which features a deeper architecture that can capture richer and more abstract feature representations. This design gives TernaNet-16 stronger feature extraction and transfer learning capabilities, as VGG16 is typically pretrained on large-scale image classification datasets, allowing it to learn rich, general-purpose image features. Additionally, TernaNet-16 cleverly employs skip connections and deconvolution operations to enhance feature fusion between the encoder and decoder while retaining U-Net's advantages in image segmentation. To further boost the model's nonlinear capabilities, TernaNet-16 adopts batch normalization and Rectified Linear Unit (ReLU) activation functions, thereby improving the model's expressiveness and generalization capabilities.

By utilizing foreground segmentation, we eliminated the interference caused by image appearance variations. We evaluated the performance of the segmentation module using the metrics Dice coefficient and Jaccard index to ensure that it effectively provided clear and accurate input for subsequent pose estimation tasks. Our experiments on real surgical scene data show that our method stably and robustly estimates the pose of surgical instruments, even in the presence of partial errors in the segmentation results.

In the main contour-based pose estimation network, we adopted a two-stage framework to estimate instrument pose from surgical instrument foreground mask images. At the first stage, our main framework estimated the 2D projection points of the implicit model feature points, and, at the second stage, the pose was predicted from the projection points and the contour points of the foreground mask.

2.3. Pose Estimation Module

2.3.1. Multi-Scale Heatmap Feature Generation

The first part of our network is a feature extraction net, a multi-scale heatmap feature module, which consists of two sub-modules: a multi-resolution convolution module and an attention (CSTSA) net, inspired by a polarized self-attention network [43], as shown in Figure 4.

A multi-resolution convolution module is a four-stage cascaded convolution network that generates both high-resolution features and low-resolution features. The first stage consists of a high-resolution convolution branch, one-fourth of the original image resolution. Then, the following three stages gradually expand parallel branches, with the resolution of each new expanded branch being half of the lowest resolution of the previous stage. Then,

the features of different resolutions are fused into a unified feature before being inputted to the next stage to preserve both the high-resolution local features and the low-resolution global features. After the last multi-resolution convolution stage, a feature map is generated by a Softmax layer.

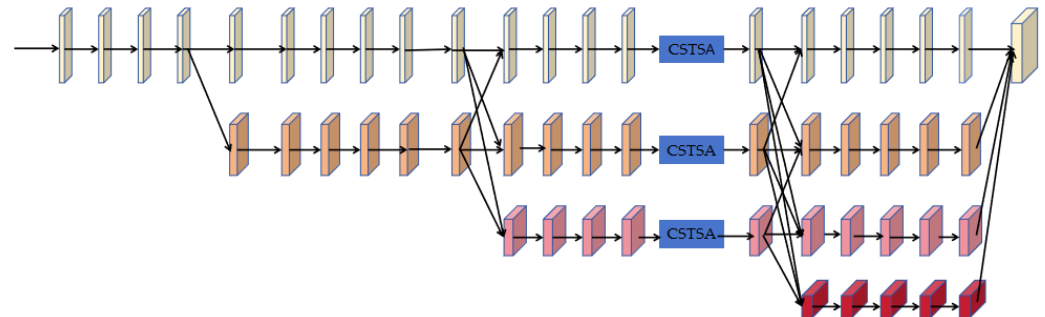


Figure 4. The multi-scale heatmap feature module to generate the feature map. The colors of the blocks indicate the feature map size. The yellow feature map is 1/4 the size of the original image, with 32 channels. The orange feature is 1/8 the size of the original image, with 64 channels. The pink feature is 1/16 the size of the original image, with 128 channels. The red feature is 1/32 the size of the original image, with 256 channels.

To fully utilize the dependency information of different parts of the intermediate features after the third multi-resolution convolution stage, we leverage channel-space tensor self-attention network. Channel-space tensor self-attention network treats the intermediate features as a tensor and extracts attention from channel-U and channel-V attention and space only attention. Channel-U and channel-V attentions are the attentions acting on the space spanned by a channel direction vector and one base vector of space. The architecture of the CSTSA module is shown in Figure 5.

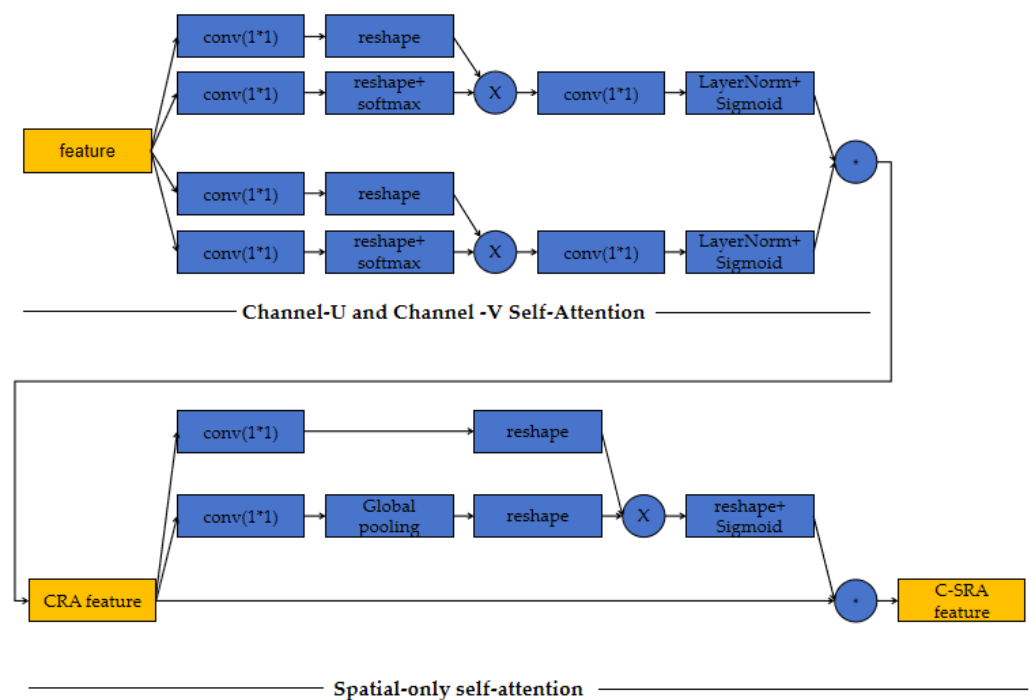


Figure 5. Architecture of the CSTSA module.

The two attention sub-modules extract the dependencies between space and channel, ensuring our model generates heatmap features that contain rich multi-scale information

and fuse spatial and channel correlation information. We set the heatmap loss to supervise our heatmap as follows:

$$L_h = \text{MSE}(m, m^*) \quad (1)$$

$\text{MSE}(\cdot, \cdot)$ is the mean squared error function in Equation (1), m is the estimated feature point heatmap, and m^* is the ground truth feature point heatmap. With our synthetic dataset, as the ground truths of the projections of implicit feature points are known, the ground truth feature point heatmap is generated by setting impulses at all the points' ground truth positions and convolving them with a Gaussian kernel with the following parameter setting: $\sigma = 8$ pixels.

2.3.2. Implicit Feature Point Projection Regression

After the extraction of feature heatmaps, we utilized a simple implicit feature point projection regression module to predict the coordinates. This module is a differentiable network to regress the pixel coordinates from the heatmap, and its architecture is shown in Figure 6.

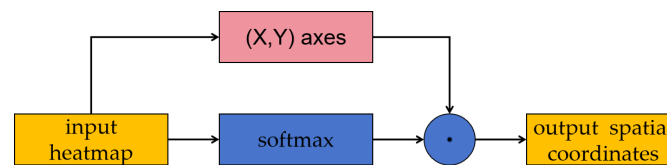


Figure 6. Architecture of implicit feature point projection regression module.

2.3.3. Surgical Instrument Pose Estimation

After the inference of the projected implicit feature points, we utilized the back-propagatable PnP(BPnP) to estimate the pose of the surgical instrument by solving the originally non-differentiable implicit function through gradient-based back-propagation. The BPnP loss function is the sum of the reprojection errors of the implicit feature points, as shown in Equation (2).

$$\begin{aligned} l_p(x, x^*, y, z, K) &= \sum_{i=1}^n \|r_i\|_2^2 + \lambda \|r_i^*\|_2^2 \\ &= \sum_{i=1}^n \|x_i - (y, z_i, K)\|_2^2 + \lambda \sum_{i=1}^n \|x_i^* - (y, z_i, K)\|_2^2 \end{aligned} \quad (2)$$

where x is the inferred projections of implicit feature points, x^* is the ground truth of the projections, y is the pose parameters, z is the 3D implicit feature points, K is the intrinsic matrix of the camera, r_i is the reprojection error of the implicit feature point, and π is the projection function of a 3D point to the image one.

We leverage the zero-point theorem to convert the task of solving pose parameters by minimizing the reprojection loss into solving the explicit equation derived from partial differentials of the reprojection loss with respect to the pose parameters, as shown in Equation (3).

$$f(x, y, z, K) = \frac{\partial l_p(x, y, z, K)}{\partial y} = 0 \quad (3)$$

According to the implicit function theorem, the partial differential of the pose parameters with respect to the reprojection errors of the implicit feature point coordinates can be computed as follows:

$$\frac{\partial y}{\partial x} = -[\frac{\partial f}{\partial y}]^{-1} [\frac{\partial f}{\partial x}] \quad (4)$$

Thus, the explicit expression of the gradient of reprojection loss with respect to the network parameters θ is as follows:

$$\frac{dl_p}{d\theta} = \frac{\partial l_p}{\partial y} \frac{\partial y}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial l_p}{\partial x} \frac{\partial x}{\partial \theta} \quad (5)$$

By minimizing the final loss function, this gradient is back-propagated in end-to-end to supervise the network parameters of the multi-scale heatmap feature module and the implicit feature point projection regression module, and finally, the surgical instrument pose is estimated.

To fully use the mask contour information, we also set a differentiable loss function to enhance the contour global consistency; the foreground rendering loss is defined as follows:

$$L_r = \text{MSE}(m_r, m_s) \quad (6)$$

where m_r is the rendered foreground image, and m_s is the segmented foreground image. We set the final loss as the weighted sum of the reprojection loss, the foreground consistency loss, and the heatmap loss, as follows:

$$l_{\text{final}} = l_h + \alpha l_p + \beta l_r \quad (7)$$

By minimizing this final loss, we trained our surgical instrument pose estimation in a weakly supervised manner. Our model utilizes synthetic foreground mask image data generated by a virtual engine as the training set and tackles the problems of stable tracking and lack of pose annotation in weakly supervised learning. In a real-data experiment, our method must be validated to be robust in estimating the surgical instrument pose, even for instruments with slightly different 3D shapes. In Equation (7), the hyper parameter α is set to 0.00005, and β is set to 1.2.

3. Results

To verify the validity of our method, experiments on both real surgery scenes and synthetic data were carried out. For real surgery scene experiments, we tested our method on our self-collected minimally invasive lung lobe resection endoscopic surgery dataset in collaboration with the Renmin Hospital affiliated with Peking University. For the synthetic data experiments, we tested our method on the UE 5 virtual engine generated instrument data. We undertook comparisons between the results of our method and the current state-of-the-art single-view Red Green Blue (RGB) image-based object pose estimation techniques on various popular metrics. To evaluate the impact of each submodule on our method, we conducted an ablation study on both real surgery scene data and synthetic data.

3.1. Metrics

The commonly used metrics of Average Distance of 3D points (ADD)(S)(0.1 d) [44,45], 2D projection error(5 px) [46], and 2D projection error(3 px) were used for the synthetic data experiments. The intersection over union (IOU) was used for real-scene data experiments. ADD refers to the average distance between the 3D model points obtained using ground truth poses and predicted poses. Typically, if the average distance of model points is less than 10% of the model diameter, the pose estimation is considered correct. ADD-S is a specific variant used for measuring symmetric objects. The 2D projection error refers to the average distance between the 2D feature points obtained from the ground truth pose projection and that from the predicted pose projection. The definitions of these metrics are shown as Equations (8)–(11).

$$\text{ADD} = \frac{1}{n} \sum_{i=1}^n \| (R x_i + t) - (\bar{R} x_i + \bar{t}) \| \quad (8)$$

$$ADD - S = \frac{1}{n} \sum_{i=1}^n \min_{x_j} \|(Rx_i + t) - (\bar{R}x_j + \bar{t})\| \quad (9)$$

$$2D \text{ projection error} = \frac{1}{m} \sum_{i=1}^m \|K(Rx_i + t) - K(\bar{R}x_i + \bar{t})\| \quad (10)$$

$$IOU = \frac{|A \cap B|}{|A \cup B|} \quad (11)$$

where x_i is the three-dimensional model point of the object, R is the predicted rotation matrix, t is the predicted translation vector, K is the camera intrinsic matrix, $\bar{R}$ is the ground truth of the rotation matrix, $\bar{t}$ is the ground truth of the translation vector, and x_j is the nearest distance point in the corresponding points of the symmetric object. IOU: A and B represent the foreground masks of instrument reprojection and segmentation.

3.2. Implementation Detail

We scaled the input image to a 640×480 resolution for both training and testing, as in [47]. We trained the network for 130 epochs. The initial learning rate was set to 2×10^5 . After 50 epochs, it was adjusted to 5×10^6 . After 100 epochs, it was further adjusted to 2×10^6 . After 120 epochs, it was set to 2×10^7 . To accelerate iterative training, the first 50 epochs were trained using only heatmap loss. The following 70 epochs were trained using a combination of heatmap loss and the reprojection error of the implicit feature points. The final 10 epochs were trained using the final loss. We used Adam as our optimizer with a momentum of 0.9 and a weight decay of 1×10^3 .

The experimental results are reported as follows:

3.3. Comparisons of Pose Estimation Results

3.3.1. Real Surgery Scene Data Experiments

In our self-collected lung endoscopic surgery dataset, including 2000 image frames extracted from real surgical video recordings, we predicted the pose of three kinds of instruments: hook, suction, and clippers, using our method trained on our synthetic dataset of 50,000 pose labeled images. Figure 7 illustrates some examples of the reprojected masks of the three kinds of instruments under the pose inferred by our method. In Figure 7, the blue points on the mask are the estimated feature points, and these images show that the reprojected masks are closely overlaid on the instrument, which indicates that the estimated poses are accurate.

As pose ground truth is unavailable for real-scene data, we compared our method with state-of-the-art methods on the IOU metric. The comparison methods include the following: Segmentation-driven-pose [20], Transpose + RansacPnP [48], Higherhrnet + RansacPnP [49], and PVNet [26]; the results are shown in Table 1. The results show that our method outperforms the others by a significant margin.

Table 1. Comparison of our method with state-of-the-art methods on the IOU metric.

Model	IOU			
	Sucker	Hook	Clipper	Average
Segmentation-driven-pose	48.20	42.91	26.42	39.18
Transpose + RansacPnP	41.22	43.68	41.58	42.16
Higherhrnet + RansacPnP	64.65	42.06	51.84	52.85
PVNet	59.24	33.67	45.59	46.17
Ours	67.06	57.71	52.89	59.22

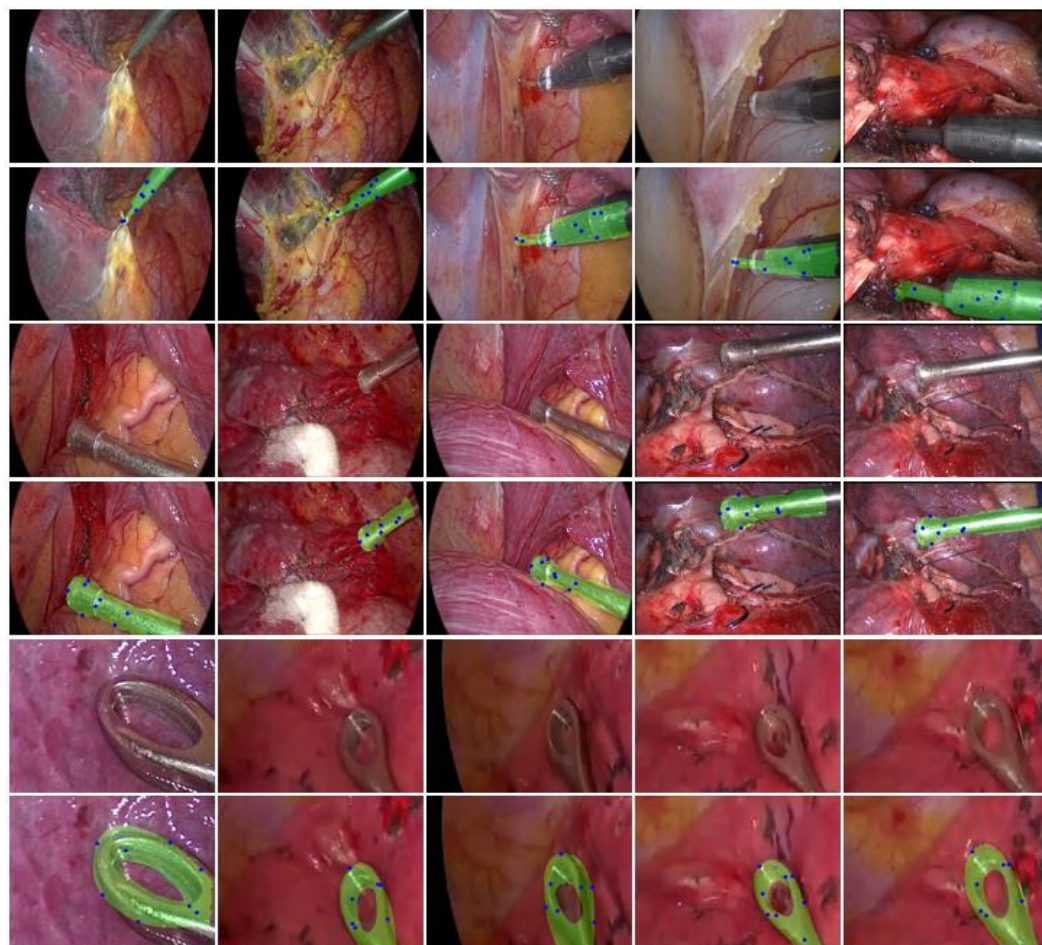


Figure 7. Examples of the reprojected masks of hook, suction, and clippers under the pose inferred by our method.

3.3.2. Synthetic Data Experiments

In our synthetic dataset experiments, we compared our methods with those methods on ADD(S)(0.1 d), 2D projection error(5 px), and 2D projection error(3 px) as the synthetic images contain the real pose information. Our dataset contains 50,000 images of 3 kinds of instruments in different poses (roughly the same number of images for each instrument). We train the compared models with 90% of the data from the dataset and the test models with the remaining 10% of the data. The results of the comparisons are shown in Table 2. The results show that the pose estimated by our method is significantly more accurate on all the three metrics.

Table 2. Comparison of our method with state-of-the-art methods, on the ADD(S)(0.1 d), 2D projection(5 px), and (3 px) metrics.

Model	ADD(S)(0.1 d)				2D Projection Error(5 px)				2D Projection Error(3 px)			
	Sucker	Hook	Clipper	Average	Sucker	Hook	Clipper	Average	Sucker	Hook	Clipper	Average
Segmentation-driven-pose	98.48	63.63	98.48	86.86	28.79	42.42	10.61	27.27	7.58	16.67	4.55	9.60
Transpose + RansacPnP	48.48	48.48	75.76	57.57	6.06	53.03	43.94	34.34	1.52	13.64	12.12	9.09
Higherhrnet + RansacPnP	100	77.27	100	92.42	37.88	86.36	69.70	64.65	10.61	56.06	34.85	33.84
PVNet	98.48	33.33	100	77.27	37.88	28.79	80.30	48.99	12.12	10.61	42.42	21.72
Ours	100	86.36	100	95.45	46.97	89.39	84.85	73.74	19.70	68.18	43.94	43.94

3.4. Ablation Study

In order to investigate the effects of each module of our method, we conducted an ablation study with both real-scene and synthetic data. We employed various combinations of multi-resolution convolution (MrC) modules, channel-space tensor self-attention (CSTSA) modules, BPnP modules, and differentiable rendering pose estimation (DRPE) modules to assess the impact of different modules, and the results are shown in Tables 3 and 4.

Table 3. Comparison of different combinations of our modules on the ADD(S)(0.1 d), 2D projection(5 px), and (3 px) metrics, with synthetic data.

Model	Synthetic Data											
	ADD(S)(0.1 d)				2D Projection Error(5 px)				2D Projection Error(3 px)			
	Sucker	Hook	Clipper	Average	Sucker	Hook	Clipper	Average	Sucker	Hook	Clipper	Average
MrC	98.48	82.83	99.49	93.6	33.84	86.36	79.29	66.50	10.61	56.06	33.84	33.50
MrC + CSTSA	100	79.80	100	93.26	43.43	86.87	78.28	69.53	16.67	63.64	38.89	39.73
MrC + CSTSA + BPnP	99.49	86.36	100	95.28	44.44	86.36	81.82	70.87	17.68	65.15	44.44	42.42
MrC + CSTSA + BPnP + DRPE	100	86.36	100	95.45	46.97	89.39	84.85	73.74	19.70	68.18	43.94	43.94

Table 4. Comparison of different combinations of our modules on the IOU metric with real data.

Model	IOU			
	Sucker	Hook	Clipper	Average
MrC	61.32	41.02	31.31	44.55
MrC + CSTSA	63.56	46.09	30.17	46.61
MrC + CSTSA + BPnP	64.78	44.99	33.94	47.90
MrC + CSTSA + BPnP + DRPE	67.06	57.71	52.89	59.22

The above ablation study shows that each module has a positive impact on the results. More specifically, the DRPE module has made the most contribution to the IOU metric in real experiments. For synthetic data, all our modules have positive contributions to the ADD(S) metric, and DRPE and CSTSA have the greatest impact on the 2D projection metric.

4. Conclusions and Discussion

In this work, we have proposed a new weakly supervised method for 6D pose estimation of surgical instruments from single endoscopic images. Our method requires only virtual engine generated synthetic instrument foreground masks as training data, without extensive manually labeled effort. Our method has successfully addressed the common challenges of stable feature tracking and the lack of labeled data in real surgical instrument pose estimation. Experiments on both self-collected real surgical instrument datasets and synthetic data demonstrate that, compared to state-of-the-art methods, our method exhibits enhanced robustness and reliability in pose estimation tasks. Our work offers a new solution for surgical instrument pose estimation that is applicable not only to surgical instrument pose estimation tasks but also provides valuable insights for weakly supervised object pose estimation problems in other domains.

Since our method uses mask images for pose estimation, our method is less affected by the illumination changes and lack of salient textures on surgical instruments. However, the contour-based methods also have their inherent weaknesses. For example, how do you choose the implicit feature points on the instrument so that they are visible under most endoscope poses? If some of the model points are invisible, whether a visible subset of feature points could still reliably predict the pose, etc.

During the experiments, we observed that in some cases, only five or six feature points out of the total eight were visible, and their corresponding estimated poses were still accurate enough. We thought that this was largely aided by our global contour consistency enforcement since the contour shape itself already imposes some effective constraints on the possible pose space.

In the future, we would like to investigate the reliability of our method on other surgical instruments, as any real surgical operation usually involves a variety of tools. In addition, the current work is based on a single endoscopic image, in real surgery, video sequences are available, and the time domain continuity will be explored to further boost the pose estimation's robustness and accuracy.

Author Contributions: Methodology, L.H., S.F. and B.W.; software, S.F., writing, B.W., S.F. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62273346).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Peking University People's Hospital (Project identification code 2022PHB011-002).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Our synthetic surgical instrument foreground mask training dataset and real surgical scene test dataset will be shared on our website if the paper is accepted.

Acknowledgments: The authors thank the Peking University People's Hospital for providing the minimally invasive lung lobe resection endoscopic surgery scene data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Garrow, C.R.; Kowalewski, K.F.; Li, L.; Wagner, M.; Schmidt, M.W.; Engelhardt, S.; Hashimoto, D.A.; Kenngott, H.G.; Bodenstedt, S.; Speidel, S.; et al. Machine learning for surgical phase recognition: A systematic review. *Ann. Surg.* **2021**, *273*, 684–693. [CrossRef] [PubMed]
2. Kawka, M.; Gall, T.M.H.; Fang, C.; Liu, R.; Jiao, L.R. Intraoperative video analysis and machine learning models will change the future of surgical training. *Intell. Surg.* **2022**, *1*, 13–15. [CrossRef]
3. DaVinci Intuitive Surgical—Procedures. Available online: <https://www.intuitive.com/en-us/products-and-services/da-vinci/education> (accessed on 25 March 2023).
4. D'Ettorre, C.; Mariani, A.; Stilli, A.; Baena, F.R.Y.; Valdastrì, P.; Deguet, A.; Kazanzides, P.; Taylor, R.H.; Fischer, G.S.; DiMaio, S.P.; et al. Accelerating surgical robotics research: A review of 10 years with the da Vinci research kit. *IEEE Robot. Autom. Mag.* **2021**, *28*, 56–78. [CrossRef]
5. Lajkó, G.; Nagyne Elek, R.; Haidegger, T. Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. *Sensors* **2021**, *21*, 5412. [CrossRef]
6. Attanasio, A.; Alberti, C.; Scaglioni, B.; Marahrens, N.; Frangi, A.F.; Leonetti, M.; Biyani, C.S.; De Momi, E.; Valdastrì, P. A comparative study of spatio-temporal U-nets for tissue segmentation in surgical robotics. *IEEE Trans. Med. Robot. Bionics* **2021**, *3*, 53–63. [CrossRef]
7. Kitaguchi, D.; Takeshita, N.; Matsuzaki, H.; Takano, H.; Owada, Y.; Enomoto, T.; Oda, T.; Miura, H.; Yamanashi, T.; Watanabe, M.; et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg. Endosc.* **2020**, *34*, 4924–4931. [CrossRef] [PubMed]
8. Nguyen, X.A.; Ljuhar, D.; Pacilli, M.; Takano, H.; Owada, Y.; Enomoto, T.; Oda, T.; Miura, H.; Yamanashi, T.; Watanabe, M.; et al. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Comput. Methods Programs Biomed.* **2019**, *177*, 1–8. [CrossRef] [PubMed]
9. Gao, Y.; Vedula, S.S.; Reiley, C.E.; Ahmidi, N.; Varadarajan, B.; Lin, H.C.; Tao, L.; Zappella, L.; Bejar, B.; Yuh, D.D.; et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In Proceedings of the MICCAI Workshop: M2cai, Cambridge, MA, USA, 14–18 September 2014; Volume 3, p. 3.
10. Pan, M.; Wang, S.; Li, J.; Li, J.; Yang, X.; Liang, K. An automated skill assessment framework based on visual motion signals and a deep neural network in robot-assisted minimally invasive surgery. *Sensors* **2023**, *23*, 4496. [CrossRef] [PubMed]
11. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VI 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 467–483.

12. Truong, P.; Apostolopoulos, S.; Mosinska, A.; Stucky, S.; Ciller, C.; Zanet, S.D. Glampoints: Greedily learned accurate match points. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10732–10741.
13. Rad, M.; Lepetit, V. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 33828–33836.
14. Tekin, B.; Sinha, S.N.; Fua, P. Real-time seamless single shot 6d object pose prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 292–301.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
16. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An accurate O (n) solution to the PnP problem. *Int. J. Comput. Vis.* **2009**, *81*, 155–166. [[CrossRef](#)]
17. Gupta, K.; Petersson, L.; Hartley, R. Cullnet: Calibrated and pose aware confidence scores for object pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Republic of Korea, 27–28 October 2019.
18. Zhao, Z.; Peng, G.; Wang, H.; Fang, H.-S.; Li, C.; Lu, C. Estimating 6D pose from localizing designated surface keypoints. *arXiv* **2018**, arXiv:1812.01387.
19. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 483–499.
20. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-driven 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3385–3394.
21. Park, K.; Patten, T.; Vincze, M. Pix2Pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 7 October–2 November 2019; pp. 7668–7677.
22. Zakharov, S.; Shugurov, I.; Ilic, S. DPOD: 6d pose object detector and refiner. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1941–1950.
23. Li, Z.; Wang, G.; Ji, X. CDPN: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7678–7687.
24. Lu, Z.; Zhang, Y.; Doherty, K.; Severinsen, O.; Yang, E.; Leonard, J. SLAM-supported self-training for 6D object pose estimation. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; IEEE: New York, NY, USA, 2022; pp. 2833–2840.
25. Weng, Y.; Wang, H.; Zhou, Q.; Qin, Y.; Duan, Y.; Fan, Q.; Chen, B.; Su, H.; Guibas, L.J. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13209–13218.
26. Peng, S.; Liu, Y.; Huang, Q.; Bao, H. PVNet: Pixel-wise voting network for 6dof pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4561–4570.
27. Song, C.; Song, J.; Huang, Q. HybridPose: 6d object pose estimation under hybrid representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 431–440.
28. He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; Sun, J. PVN3D: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11632–11641.
29. Do, T.T.; Cai, M.; Pham, T.; Reid, I. Deep-6dpose: Recovering 6d object pose from a single rgb image. *arXiv* **2018**, arXiv:1802.10367.
30. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
31. Liu, F.; Fang, P.; Yao, Z.; Fan, R.; Pan, Z.; Sheng, W.; Yang, H. Recovering 6D object pose from RGB indoor image based on two-stage detection network with multi-task loss. *Neurocomputing* **2019**, *337*, 15–23. [[CrossRef](#)]
32. Girshick, R. Fast R-CNN. In Proceedings of the IEEE international Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
33. Sundermeyer, M.; Marton, Z.C.; Durner, M.; Brucker, M. Implicit 3d orientation learning for 6d object detection from rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 699–715.
34. Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; Fox, D. DeepIM: Deep iterative matching for 6d pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 683–698.
35. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
36. Deng, X.; Xiang, Y.; Mousavian, A.; Eppner, C.; Bretl, T.; Fox, D. Self-supervised 6d object pose estimation for robot manipulation. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May 2020; pp. 3665–3671.

37. Wang, G.; Manhardt, F.; Shao, J.; Ji, X.; Navab, N.; Tombari, F. Self6D: Self-supervised monocular 6d object pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16. Springer International Publishing: Cham, Switzerland, 2020; pp. 108–125.
38. Chen, B.; Parra, A.; Cao, J.; Li, N.; Chin, T.J. End-to-end learnable geometric vision by backpropagating pnp optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8100–8109.
39. Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.Y.; Johnson, J.; Gkioxari, G. Accelerating 3d deep learning with pytorch3d. *arXiv* **2020**, arXiv:2007.08501.
40. Shvets, A.A.; Rakhlin, A.; Kalinin, A.A.; Iglovikov, V.I. Automatic instrument segmentation in robot-assisted surgery using deep learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 624–628.
41. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv* **2021**, arXiv:2107.00782.
44. Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In Proceedings of the Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Republic of Korea, 5–9 November 2012; Revised Selected Papers, Part I 11; Springer: Berlin/Heidelberg, Germany, 2013; pp. 548–562.
45. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv* **2017**, arXiv:1711.00199.
46. Brachmann, E.; Michel, F.; Krull, A.; Yang, M.Y.; Gumhold, S. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3364–3372.
47. Redmon, J.; Farhadi, A. YOLOV3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
48. Yang, S.; Quan, Z.; Nie, M.; Yang, W. Transpose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11802–11812.
49. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5386–5395.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.