

# SurgeoNet: Realtime 3D Pose Estimation of Articulated Surgical Instruments from Stereo Images using a Synthetically-trained Network

Ahmed Tawfik Aboukhadra<sup>1,2</sup>, Nadia Robertini<sup>1</sup>, Jameel Malik<sup>3</sup>, Ahmed Elhayek<sup>4</sup>, Gerd Reis<sup>1</sup>, and Didier Stricker<sup>1,2</sup>

<sup>1</sup> German Research Center for Artificial Intelligence (DFKI), Trippstadter Straße 122, 67663 Kaiserslautern, Germany

{firstname[\_secondname].lastname}@dfki.de

<sup>2</sup> University of Kaiserslautern-Landau (RPTU), Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany

<sup>3</sup> NUST-SEECS, Islamabad, Pakistan

<sup>4</sup> University of Prince Mugrin (UPM), 42241 Madinah, Saudi Arabia

**Abstract.** Surgery monitoring in Mixed Reality (MR) environments has recently received substantial focus due to its importance in image-based decisions, skill assessment, and robot-assisted surgery. Tracking hands and articulated surgical instruments is crucial for the success of these applications. Due to the lack of annotated datasets and the complexity of the task, only a few works have addressed this problem. In this work, we present SurgeoNet, a real-time neural network pipeline to accurately detect and track surgical instruments from a stereo VR view. Our multi-stage approach is inspired by state-of-the-art neural-network architectural design, like YOLO and Transformers. We demonstrate the generalization capabilities of SurgeoNet in challenging real-world scenarios, achieved solely through training on synthetic data. The approach can be easily extended to any new set of articulated surgical instruments. SurgeoNet’s code and data are publicly available<sup>5</sup>.

**Keywords:** Mixed Reality · Computer Vision · Deep Learning · Object Detection · 3D Pose Estimation · Transformer

## 1 Introduction

Recent advancements in virtual and augmented reality technology have enabled highly immersive gaming, interactive simulation, and virtual experiences, among others. The use of mixed reality in the medical field is gaining attention, especially in the simulation of medical scenarios for the training of medical personnel outside the laboratory. To ensure a highly immersive experience, attention goes into realistic user interactions with virtual objects. This includes realistic object modeling and pose tracking at low latency. The vast majority of the existing

<sup>5</sup> <https://github.com/ATAboukhadra/SurgeoNet>

approaches in the field of object-tracking, mainly focus on large rigid objects of everyday use, such as books, cans, and cups [5,6,1]. The problem of tracking semi-rigid objects is understudied and has only recently gained attention [4,26,4,11]. Especially in the case of tracking surgical instruments, to the best of our knowledge, there exists little to no advance, due to its intrinsic difficulties [21,8]. Surgical semi-rigid instruments, including thin scissors of various types and forceps or clamps, have strong similarities in shape and appearance. On top of it, hand interaction reduces their visibility, causing typical Computer Vision solutions to struggle to identify or classify them. General neural-network-based solutions require tons of finely labeled data for training to succeed at the task. However, such data is hard to obtain automatically.

In this work, we introduce SurgeoNet, a real-time solution to the problem of surgical instrument tracking from a stereo (VR) view, that does not rely on a realistic dataset. Instead, our approach builds on top of synthetically generated samples of surgical instruments.

We design our pipeline to infer real-time 7D surgical instruments' pose (3D translation, + 3D rotation + 1D articulation angle) from stereo view and demonstrate its capabilities in detecting, classifying, and tracking instruments in real settings as seen from VR glasses. Our method consists of two main components: the first is designed for object detection, classification, and 2D keypoint estimation. The second and final part of the pipeline combines the keypoints obtained from left and right stereo-view to infer the corresponding pose. Our method is robust to occlusions due to hand interactions and accurately classifies surgical instruments, despite their strong similarities. Thanks to its real-time computational performances, our method is suitable for virtual and augmented reality applications, enabling realistic and highly immersive interactions with realistic virtual medical tools. The method is easily extendable to a different subset of rigid or semi-rigid medical instruments, currently with at most 1 degree of freedom, and the introduction of new objects in the set is straightforward.

In summary, this work presents SurgeoNet, a new method to accurately reconstruct 7D poses of articulated surgical instruments from stereo view, with the following key features:

1. Real-time performance, thus suitable for mixed-reality applications;
2. Reliable classification of surgical instruments of similar shape and appearance under occlusions;
3. Temporally consistent, jitter-free, tracking.
4. High generalization capabilities to unseen (real) sequences, despite relying solely on a synthetic dataset.

## 2 Related Work

In our review of related works, we subdivide our problem into three fields: surgical instrument pose estimation, stereo-based object pose estimation, and articulated object pose estimation.

Multiple works studied the surgical instruments’ pose estimation problem [21,8]. Rodrigues *et al.* [17] published a survey of all datasets of surgical instruments. Most of those datasets, however, only contain 2D annotations i.e. instrument labels, bounding boxes, 2D keypoints, or at best, segmentation masks. An example of recent surgical instruments datasets that contain 2D labels and keypoints is PWISeg [18]. Hein *et al.* [8] proposed a clinical dataset that includes synthetic and real monocular RGB images for hands interacting with a surgical drill along with the 6D annotations of the instrument. They also propose a pipeline for hand-object pose estimation, however, it’s only focused on a single fully-rigid object i.e. drill. In their experiments, they compare the performance of PVNet [14] and HandObjectNet [7]. HMD-EgoPose [3] uses an EfficientNet [19] as a backbone to predict the drill pose in the Hein *et al.* dataset. In addition, the authors deploy their method on a Microsoft HoloLens 2 AR headset. In our experiments section, we finetune our network on the Hein *et al.* real dataset and report pose estimation errors.

POV-Surgery [21] is another work that provides a synthetic dataset that considers temporal dependencies. It includes hands wearing stained surgical gloves and interacting with surgical instruments. The authors provide a finetuned hand pose estimation model to handle those special-looking hands. POV-Surgery focuses only on a small set of completely rigid surgical instruments.

Given our interest in stereo-based vision, we also study the stereo-RGB methods for object pose estimation. StereOBJ-1M [12] is a large-scale dataset that contains stereo RGB frames of 18 objects and their 3D pose annotations. KeyPose [13] is one of the famous methods meant for stereo-based pose estimation and was evaluated on the StereOBJ-1M dataset. KeyPose is a neural network that predicts 3D keypoints of rigid transparent objects from stereo input. One of the key ideas in KeyPose is that they use early fusion in their CNN-based network. This means that features from left and right views are merged earlier in the pipeline which improves performance.

Recently, more attention has been given to articulated object pose estimation [26,4,11]. The ARCTIC dataset [4] provides a real RGB dataset with full annotations of hands dexterously manipulating articulated objects. However, the set of objects provided in those datasets doesn’t include surgical instruments and only focuses on everyday objects with varying textures and shapes.

### 3 Method

Given a sequence of calibrated stereo pairs of RGB images, captured from typical VR glasses, mimicking a user’s eyes, SurgeoNet estimates the 7D pose of the visible surgical instruments in the camera coordinate system. We propose a neural network pipeline consisting of three stages: 1. Object detection and keypoints estimation. 2. Tracking and temporal smoothing of keypoints and labels. 3. 7D Pose Estimation from keypoints and labels. Figure 1 shows our selected architecture. In this section, we describe each network component as well as the synthetic dataset generation process in detail.

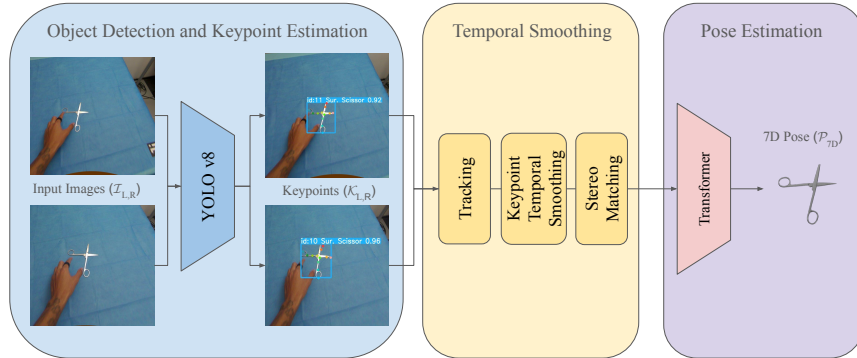


Fig. 1: SurgeoNet Architecture.

**Object Detection and Keypoints Estimation** The first component of the pipeline detects the surgical instruments present in the views and estimates the corresponding keypoints  $\hat{\mathcal{K}}_{2D,c}$ . This stage is implemented using an enhanced version of the YOLO [16] architecture, namely YOLOv8 [9]. YOLOv8 contains advanced backbone and neck architectures for improved feature extraction. All YOLO’s predictions with a confidence value below 0.7 are discarded.

**Tracking and temporal smoothing of keypoints and labels** We use ByteTrack [22], a state-of-the-art tracker, to assign a tracking ID to the detected bounding boxes, taking into account their historical positions. ByteTrack is a MOT algorithm that improves over previous MOT algorithms by associating almost all detections instead of only the high-scoring ones which improves the tracking in case of occlusions. ByteTrack predicts the new location of tracks from previous frames using Kalman Filters. It then uses similarity metrics like IoU and Re-ID to associate new bounding boxes to the tracks. This way it can keep unique IDs for the tracks. Those IDs are used later to apply temporal smoothing to keypoints and reduce jitter. For this task we use 1€ Filter [2]. Temporally smoothed keypoints and bounding boxes from the left and right stereo view are finally matched considering the epipolar lines.

**7D Pose Estimation from keypoints and labels** Taking inspiration from previous work [23,24], we designed a Transformer [20] network that transforms the stereo 2D keypoints denoted as  $\hat{\mathcal{K}}_{2D,c}$  of an object to its 7D Pose denoted as  $\hat{\mathcal{P}}_{7D}$  that contains 3D rotation, 3D translation, and 1 articulation value for articulated-like objects as scissors. Given that attention layers are permutation invariant i.e. the order of keypoints given to it doesn’t change its output. Therefore, we added a one-hot vector to each keypoint vector to describe its relative position to other keypoints. Furthermore, to improve the transformer’s ability

to comprehend which kind of object the keypoints describe, we also append a one-hot encoding that describes the label of the object.

The network consists of stacked multi-headed attention layers. The outputs of this network are two-fold: 1. 3D keypoints that correspond to the triangulation of the stereo keypoints  $\hat{\mathcal{K}}_{3D}$ . 2. The 7D object pose  $\hat{\mathcal{P}}_{7D}$  that contains 3 translation values, 6 rotation values as described in [25] that solves the problem of rotation continuity, and 1 value for articulation angle. To train the network, we use the synthetic dataset described in Section 3.1. After getting  $\hat{\mathcal{P}}_{7D}$  from the Transformer, it is used in  $\mathcal{T}_{wld}$  to transform mesh vertices  $\mathcal{V}_o$ . The loss function used to train this network is:

$$\mathcal{L}_p = \|\hat{\mathcal{P}}_{7D} - \mathcal{P}_{7D}\|_2 + \|\mathcal{T}_{wld}(\mathcal{V}_o, \hat{\mathcal{P}}_{7D}) - \mathcal{T}_{wld}(\mathcal{V}_o, \mathcal{P}_{7D})\|_2 + \|\hat{\mathcal{K}}_{3D} - \mathcal{T}_{wld}(\mathcal{K}_o, \mathcal{P}_{7D})\|_2 \quad (1)$$

### 3.1 Synthetic Data

**Mesh Modeling and Keypoint Selection** To generate synthetic samples of a surgical instrument, we require a 3D model of that object. This can be acquired from open-source collections or semi-automatic 3D scanning. To enable modeling of the articulation angle for some semi-rigid instruments, e.g. scissors, we manually separate the meshes into their rigid geometric components, e.g. the left and right blades. Figure 2a shows the selected set of surgical instruments. Out of those 3D models, we choose a set of keypoints that will be used later for building and training our deep neural network approach.

**Synthetic RGB Dataset** To train a network for surgical instrument detection and keypoint estimation from RGB images, we use PyTorch3D [15] to synthesize RGB images of objects in random poses. During the synthesis process, we select a random subset of the surgical instruments’ meshes and randomize their 7D pose. The corresponding rendered meshes are then projected into random backgrounds. The annotations of that image are the random 7D transformation, the 12 keypoints along with the object classes and their bounding boxes. The resulting total number of generated samples is around 10K images. Figure 2b shows samples of rendered images with plotted annotations.

**Synthetic Transformer Dataset** To train the Transformer network shown in Figure 1, we follow a similar approach to the synthetic RGB dataset in which we randomize a pose and apply it to a mesh. Instead of rendering the mesh on a random background, we project it to both left and right camera coordinates to get the stereo 2D keypoints of that pose. The final dataset has around 100k samples of stereo 2D keypoints and their corresponding 7D poses.

## 4 Experiments

The main qualitative results of our approach are summarized in Figure 3. In the following sections, we quantitatively evaluate the different components of

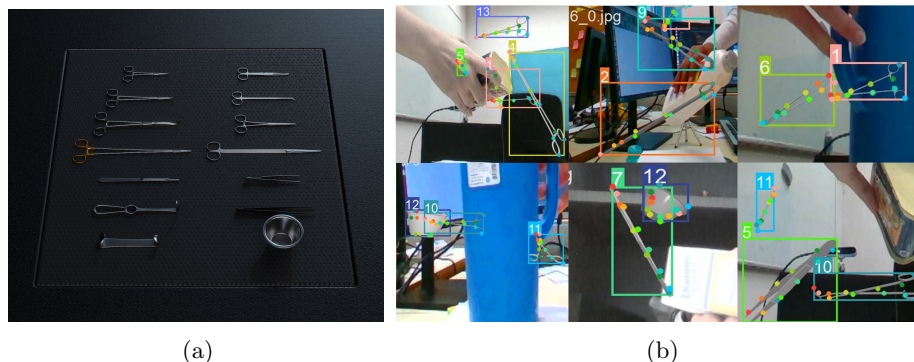


Fig. 2: a) A rendered synthetic image of the studied set of surgical instruments. b) Synthetic images generated using PyTorch3D that include medical instruments in random poses with their annotations on background images from the HO-3D dataset[5].

our pipeline. We first compare the quality and performance of different YOLOv8 architectures with varying image resolutions. We additionally show the impact of the amount of training data on the first stage and its confusion scores regarding the studied set of objects. Finally, we conduct an ablation study on the design of the Transformer network and compare it to an alternative optimization-based approach.

#### 4.1 YOLOv8 Evaluation on Surgical Instruments

**Network and Resolution** There exist multiple YOLOv8 models that differ in the number of parameters and speed. To evaluate those models on our task, we finetune them on the synthetic train dataset for 200 epochs. We then run an evaluation on the test set to measure both the bounding box and keypoint mean Average Precision (mAP) at 50-95 IoU thresholds. Runtime performance measured in Stereo Frames per Second (S-FPS) is calculated by running inference on a sequence of stereo RGB images captured using the Varjo and calculating the average over the sequence. We use the TensorRT format and run inference on an NVIDIA GeForce RTX 3090. The results are summarized in Figure 4a.

In this work, we focus on YOLOv8m with 640 resolution and YOLOv8s with 1152 resolution as they both provide real-time performance while maintaining high accuracy.

**Required amount of training samples** To evaluate the impact of the number of synthetic samples required for training the instruments' detection stage, we train two more networks on 10% and 50% of the total 8k training synthetic images. The results are summarized in Figure 4b.

**YOLOv8 Confusion Matrix** We test the classification performance of YOLO on surgical instruments and its ability to distinguish between similar objects e.g. scissors-like objects. We record a sequence for each object in our

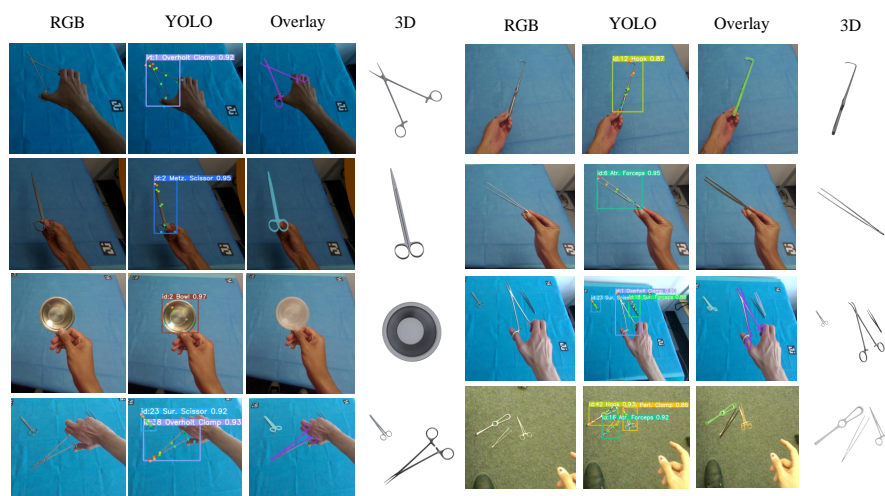


Fig. 3: The results of SurgeoNet on real unseen images.

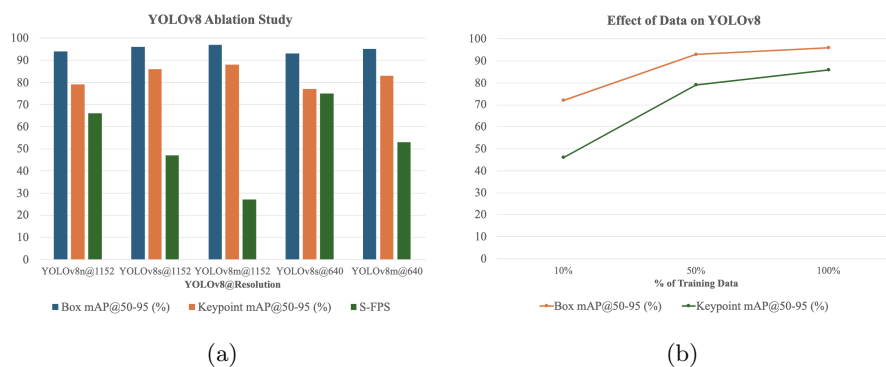


Fig. 4: YOLOv8 Ablation Study: a) The impact of the YOLOv8 architecture and image resolution on the accuracy (Box and Keypoint mAP5@50-95) and runtime performance (S-FPS). b) The impact of the amount of training data on YOLOv8’s performance.



dataset using a VR headset resulting in 13 sequences. Afterward, we run inference on all frames knowing the ground truth class of each frame, and compare it to YOLO’s predictions. Figure 5 shows the normalized confusion matrix. Except for the surgical forceps, which only differ in size, the remaining instruments are correctly classified despite their strong similarities.

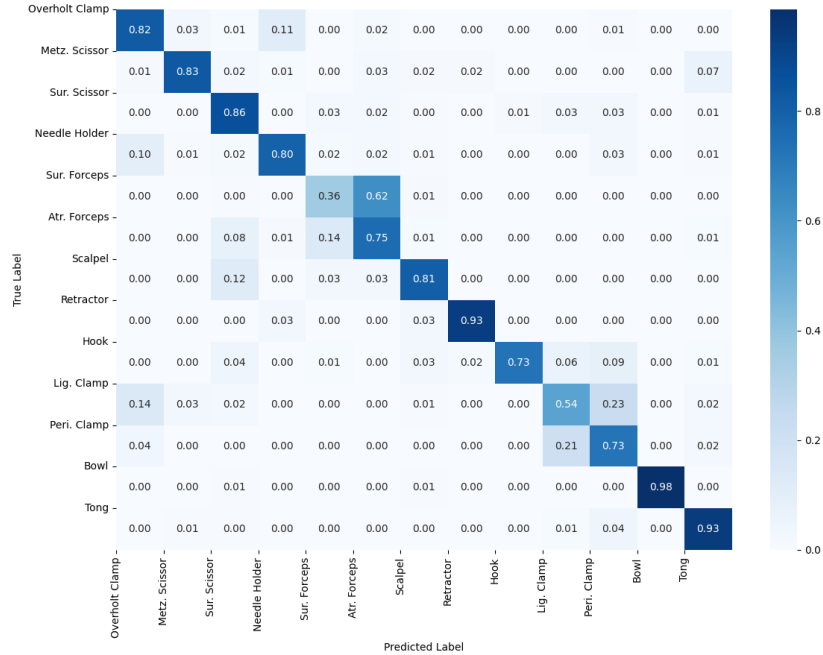


Fig. 5: Confusion matrix of YOLOv8 for the 13 surgical instruments.

## 4.2 Ablation Study on the Stereo Transformer

In this section, we show the importance of input and output formats used to design the transformer along with its hyperparameters. We show the importance of using stereo 2D keypoints instead of using only monocular 2D keypoints. Furthermore, we observe the increased performance that results from using a one-hot encoding to describe each keypoint class i.e. positional embedding for keypoints. Finally, we adopt the 6D rotation representation [25] instead of 3 axis angles. Table 1 summarizes the results of training Transformer networks on the different I/O modalities for 100 epochs and testing them on the synthetic dataset described in Section 3.1. We use the Mean Per Vertex Position Error (MPVPE) in mm to describe the quality of the pose. To study the architecture of the Transformer, we experiment with two hyperparameters, namely, the number



of multi-headed attention layers and hidden dimension representation size. No significant impact was observed by tuning the hyperparameters and the best combination for the number of layers and hidden dimension size is 5 and 128 respectively.

**Transformer vs Fitting** The final experiment is to compare the Transformer to the optimization-based keypoint fitting on two recorded real sequences.

We use an Adam optimizer [10] to optimize the 7D pose  $\hat{\mathcal{P}}_{7D}$ . During optimization,  $\mathcal{T}_{wld}$  transforms the predefined set of 3D keypoints of the object  $\mathcal{K}_o$  from their original position using the optimized pose  $\hat{\mathcal{P}}_{7D}$ . After that,  $\mathcal{T}_{cam}$  projects the mesh to the pixel coordinates of both stereo frames using the camera matrix  $\mathcal{M}_c$ . Finally, the loss is computed as the difference between the projected keypoints and  $\hat{\mathcal{K}}_{2D,c}$  predicted by YOLO.

$$\mathcal{L}_r = \frac{1}{2} \sum_c \|\mathcal{T}_{cam}(\mathcal{T}_{wld}(\mathcal{K}_o, \hat{\mathcal{P}}_{7D}), \mathcal{M}_c) - \hat{\mathcal{K}}_{2D,c}\|_2 \quad (2)$$

In the first frame of the sequence, we initialize the pose of the objects in the scene with 500 iterations of optimization. In the subsequent frames, we use the pose of the previous frame as an initialization and limit the number of iterations to 100. We also apply early stopping on the optimization process whenever the reprojection error as described in Equation 2 becomes less than 4 pixels to improve runtime while maintaining good qualitative output. From the results shown in Table 2, we can infer that the optimization process can sometimes be more accurate than the Transformer. However, optimization heavily relies on the number of iterations needed to converge which makes it very slow compared to the Transformer, and hence, not suitable for real-time applications.

Table 1: Ablation Study on the I/O Transformer Modality

Mono	Stereo	Kp Cls.	6D Rot.	MPVPE (mm)
✓	✗	✗	✗	64.0
✗	✓	✗	✗	28.9
✗	✓	✓	✗	23.0
✗	✓	✓	✓	<b>11.8</b>

Table 2: Comparison between the Transformer method and the optimization-based fitting method.

Obj. Cls.	Transformer		Optimization	
	Err.	FPS	Err.	FPS
1	16.9	209	13.8	1
13	11.8	202	21.6	1

### 4.3 State-of-the-art Comparison

To compare our method as a surgical instrument pose estimator, we train and evaluate the network on the Hein *et al.* [8] surgical drill dataset. In addition, we evaluate our method as a stereo-based object pose estimator on the StereOBJ-1M [12] benchmark dataset.

**Hein et al.** [8] (**Drill**) is a dataset containing real and synthetic monocular 256x256 frames of a surgical drill being used in an operation room. In this work, we only focus on the real dataset for training and testing. The total number of real frames is 3,746 and we follow the same fivefold cross-validation evaluation setup mentioned by the authors. To train the network, we sample 12 keypoints from the drill mesh and train a YOLOv8-m and the Transformer with the additional drill class.

Given the rigidity of the object, the ADD metric (Average Distance of Model Points) is used for evaluation:

$$\text{ADD} = \frac{1}{|M|} \sum_{x \in M} \|(Rx + t) - (\hat{R}x + \hat{t})\| \quad (3)$$

where  $M$  are the mesh 3D points,  $R$  and  $t$  represent the ground truth pose, and  $\hat{R}$  and  $\hat{t}$  represent the predicted pose. The results in Table 3 suggest that the Transformer can find an accurate pose if given correct 2D keypoints from a single view. However, when given the YOLO predictions, the network produces lower accurate poses in comparison to previous methods.

Table 3: Average ADD error across fivefold cross-validation test sets.

Model	Tool ADD (mm)
HandObjectNet [7]	13.8
PVNet [14]	39.7
HMD-EgoPose [3]	17.2
Ours (w/ perfect keypoints)	11.4
Ours (w/ YOLO keypoints)	44.3

**StereOBJ-1M** [12] contains stereo RGB frames with a resolution of 1440x1440 along with 6-DoF annotations for all rigid objects in the scene, and predefined meshes and keypoints. The total number of objects in the dataset is 18. The dataset consists of 394,612 stereo frames, of which 274,613 are used for training. To train our network, we scale the resolution of the images and keypoints to 640x640 and train a YOLOv8-m model and a Transformer for 20 epochs each.

To evaluate our method, we use the ADD-S Accuracy metric proposed by the dataset authors, defined:

$$\text{ADD-S} = \frac{1}{|M|} \sum_{x_1 \in M} \min_{x_2 \in M} \|(Rx_1 + t) - (\hat{R}x_2 + \hat{t})\| \quad (4)$$

with the same definition of variables used in Section 4.3. The ADD-S accuracy considers a pose correct if the ADD-S distance is less than 10% of the object’s diameter. Table 4 summarizes our results on the StereOBJ-1M benchmark dataset. The detailed results can be found on the challenge website<sup>6</sup>. Despite being the

lowest on average accuracy, SurgeoNet shows competitive results on multiple objects compared to State-of-the-art achieving best scores on some of them. This shows improvements on multiple objects as seen in the last row of Table 4. Figure 6 shows qualitative results on the StereOBJ-1M test set. The results suggest that our method is robust in cluttered scenes and with transparent objects.

Table 4: Average ADD-S Accuracy and ADD-S Accuracy on a selected subset of objects (abbreviated with object initials) from the StereOBJ-1M benchmark dataset.

Model	Average	M	NNP	P100	SC	STR200	TR1.5	TR50	WS
PVNet [14]	<b>42.48</b>	18.72	51.78	0.65	<b>69.11</b>	62.52	52.05	75.04	<b>72.63</b>
KeyPose [13]	39.42	<b>39.22</b>	51.72	1.77	39.12	<b>67.04</b>	60.10	72.05	71.78
Ours	36.46	29.22	<b>52.01</b>	<b>6.81</b>	51.40	59.21	<b>72.50</b>	<b>87.12</b>	62.52



Fig. 6: The results of SurgeoNet on StereOBJ-1M test set.

## 5 Conclusion

In this work, we presented SurgeoNet, a new real-time neural-network pipeline to accurately reconstruct temporally-consistent 7D poses of articulated surgical instruments from stereo VR-view. The approach builds on top of state-of-the-art architectures, including YOLO and Transformers. Thanks to its real-time performances, the approach is suitable for mixed-reality applications, especially in medical scenarios involving hand-object interactions with surgical tools. We demonstrated the method’s robustness in the classification of thin articulated

<sup>6</sup> <https://eval.ai/web/challenges/challenge-page/1645/leaderboard/3943>

surgical instruments of similar shape and appearance in challenging settings with occlusions. As shown in the evaluation, SurgeoNet demonstrated strong generalization capabilities to real sequences, despite being trained exclusively on cheap synthetic dataset.

Future work includes handling hand-object interactions to fine-tune the predicted pose. In addition, long-term temporal information from sequential frames will be used to improve the model’s performance.

**Acknowledgements:** This work was partially funded by the Federal Ministry of Education and Research of the Federal Republic of Germany (BMBF), under grant agreements: GreifbAR [Grant Nr 16SV8732], and DECODE [Grant Nr 01IW21001].

## References

1. Aboukhadra, A.T., Malik, J., Elhayek, A., Robertini, N., Stricker, D.: Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1001–1010 (2023) [2](#)
2. Casiez, G., Roussel, N., Vogel, D.: 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2527–2530 (2012) [4](#)
3. Doughty, M., Ghugre, N.R.: HMD-EgoPose: head-mounted display-based egocentric marker-less tool and hand pose estimation for augmented surgical guidance . <https://doi.org/10.1007/s11548-022-02688-y>, <https://doi.org/10.1007/s11548-022-02688-y> [3](#), [10](#)
4. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12943–12954 (2023) [2](#), [3](#)
5. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3196–3206 (2020) [2](#), [6](#)
6. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11090–11100 (2022) [2](#)
7. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Computer Vision and Pattern Recognition (CVPR) (2020) [3](#), [10](#)
8. Hein, J., Seibold, M., Bogo, F., Farshad, M., Pollefeys, M., Fürnstahl, P., Navab, N.: Towards markerless surgical tool and hand pose estimation. International journal of computer assisted radiology and surgery **16**, 799–808 (2021) [2](#), [3](#), [9](#), [10](#)
9. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), <https://github.com/ultralytics/ultralytics> [4](#)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015) [9](#)
11. Li, X., Wang, H., Yi, L., Guibas, L.J., Abbott, A.L., Song, S.: Category-level articulated object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3706–3715 (2020) [2](#), [3](#)

12. Liu, X., Iwase, S., Kitani, K.M.: Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10870–10879 (2021) [3](#), [9](#), [10](#)
13. Liu, X., Jonschkowski, R., Angelova, A., Konolige, K.: Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020) (2020) [3](#), [11](#)
14. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4561–4570 (2019) [3](#), [10](#), [11](#)
15. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d (2020) [5](#)
16. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) [4](#)
17. Rodrigues, M., Mayo, M., Patros, P.: Surgical tool datasets for machine learning research: a survey. *International Journal of Computer Vision* **130**(9), 2222–2248 (2022) [3](#)
18. Sun, Z., Xu, H., Wu, J., Chen, Z., Lei, Z., Liu, H.: Pwiseg: Point-based weakly-supervised instance segmentation for surgical instruments (2023) [3](#)
19. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020) [3](#)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [4](#)
21. Wang, R., Ktistakis, S., Zhang, S., Meboldt, M., Lohmeyer, Q.: Pov-surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 440–450. Springer (2023) [2](#), [3](#)
22. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box (2022) [4](#)
23. Zhao, W., Wang, W., Tian, Y.: Graformer: Graph-oriented transformer for 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20438–20447 (2022) [4](#)
24. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11656–11665 (2021) [4](#)
25. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019) [5](#), [8](#)
26. Zhu, Z., Wang, J., Qin, Y., Sun, D., Jampani, V., Wang, X.: Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation. arXiv preprint arXiv:2305.01618 (2023) [2](#), [3](#)