bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning

Alexey Shvets[1], Alexander Rakhlin[2], Alexandr A. Kalinin[3], and Vladimir Iglovikov[4]

[1] Massachusetts Institute of Technology, Cambridge, MA 02142, USA
shvets@mit.edu
[2] National Research University of Electronic Technology, Moscow, Russia
rakhlin@gmx.net
[3] University of Michigan, Ann Arbor, MI 48109, USA
akalinin@umich.edu
[4] Lyft Inc., San Francisco, CA 94107, USA
iglovikov@gmail.com

Abstract. Semantic segmentation of robotic instruments is an important problem for the robot-assisted surgery. One of the main challenges is to correctly detect an instrument's position for the tracking and pose estimation in the vicinity of surgical scenes. Accurate pixel-wise instrument segmentation is needed to address this challenge. In this paper we describe our winning solution for MICCAI 2017 Endoscopic Vision SubChallenge: Robotic Instrument Segmentation and its further refinement. Our approach demonstrates an improvement over the state-of-the-art results using several novel deep neural network architectures. It addressed the binary segmentation problem, where every pixel in an image is labeled as an instrument or background from the surgery video feed. In addition, we solve a multi-class segmentation problem, in which we distinguish between different instruments or different parts of an instrument from the background. In this setting, our approach outperforms other methods in every task subcategory for automatic instrument segmentation thereby providing state-of-the-art results for these problems. The source code for our solution is made publicly available at https://github.com/ternaus/robot-surgery-segmentation

Keywords: Medical imaging, Robot-assisted surgery, Computer vision, Image segmentation, Deep learning

1 Introduction

Research in robotics promises to revolutionize surgery towards safer, more consistent and minimally invasive intervention [3,14]. New developments continues on the way to robot-assisted systems and moving toward a future with fully autonomous robotic surgeons. Thus far, the most widespread surgical system is the da Vinci robot, which has already proved its favor via remote controlled laparoscopic surgery in gynecology, urology, and general surgery [3].

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Information in a surgical console of a robot-assisted surgical system includes valuable details for intra-operative guidance that can help the decision making process. This information is usually represented as 2D images or videos that contain surgical instruments and patient tissues. Understanding these data is a complex problem that involves the tracking and pose estimation for surgical instruments in the vicinity of surgical scenes. A critical component of this process is semantic segmentation of the instruments in the surgical console. Semantic segmentation of robotic instruments is a difficult task by the virtue of light changes such as shadows and specular reflections, visual occlusions such as blood and camera lens fogging, and due to the complex and dynamic nature of background tissues. Segmentation masks can be used to provide a reliable input to instrument tracking systems. Therefore, there is a compelling need for the development of accurate and robust computer vision methods for semantic segmentation of surgical instruments from operational images and video.

There is a number of vision-based methods developed for the robotic instrument detection and tracking [14]. Instrument-background segmentation can be treated as a binary or instance segmentation problem for which classical machine learning algorithms have been applied using color and/or texture features [6, 20]. Later applications addressed this problem as semantic segmentation, aiming to distinguish between different instruments or their parts [2,16].

Recently, deep learning-based approaches demonstrated performance improvements over conventional machine learning methods for many problems in biomedicine [5,12]. In the domain of medical imaging, convolutional neural networks (CNN) have been successfully used, for example, for breast cancer histology image analysis [17], bone disease prediction [21] and age assessment [10], and other problems [5]. Previous deep learning-based applications to robotic instrument segmentation have demonstrated competitive performance in binary segmentation [1, 7] and promising results in multi-class segmentation [15].

In this paper, we present a deep learning-based solution for robotic instrument semantic segmentation that achieves state-of-the-art results in both binary and multi-class setting. We used this method to produce a submission to the MICCAI 2017 Endoscopic Vision SubChallenge: Robotic Instrument Segmentation [13] that placed first, winning the competition. Here we describe the details of that solution based on a modification of the U-Net model [9,18]. Moreover, we provide further improvements over this solution utilizing recent deep architectures: TernausNet [11] and a modified LinkNet [4].

2 Methods

2.1 Dataset

The training dataset consists of 8×225-frame sequences of high resolution stereo camera images acquired from a da Vinci Xi surgical system during several different porcine procedures, see [13]. Training sequences are provided with 2 Hz frame rate to avoid redundancy. Every video sequence consists of two stereo channels

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

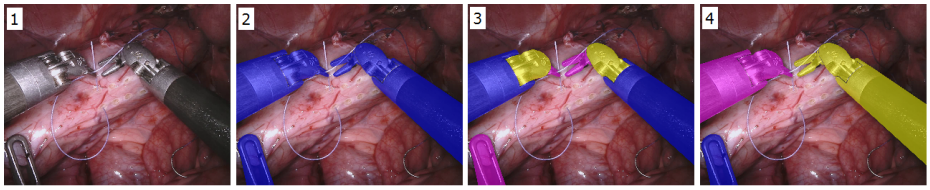

Fig. 1: A snapshot from a robotic surgical video that contains robotic instruments and patient tissues: (1) original video frame; (2) binary segmentation of robotic instruments shown in blue and tissue that serves as a background; (3) multi-class segmentation of robotic instruments where each class corresponds to a different part of the robotic instrument (3 classes: rigid shaft, articulated wrist and claspers); and (4) multi-class segmentation of robotic instruments where each class corresponds to a different robotic instrument (7 classes).

taken from left and right cameras and has a 1920×1080 pixel resolution in RGB format. To remove black canvas and extract original 1280×1024 camera images from the frames, an image has to be cropped starting from the pixel at the $(320, 28)$ position. Ground truth labels are provided for left frames only, therefore only left channel images are used for training. The articulated parts of the robotic surgical instruments, such as a rigid shaft, an articulated wrist and claspers have been hand labelled in each frame. Ground truth labels are encoded with numerical values $(10, 20, 30, 40, 0)$ and assigned to each part of an instrument or background. Furthermore, there are instrument type labels that categorize instruments in following categories: left/right prograsp forceps, monopolar curved scissors, large needle driver, and a miscellaneous category for any other surgical instruments.

The test dataset consists of 8×75-frame sequences containing footage sampled immediately after each training sequence and 2 full 300-frame sequences, sampled at the same rate as the training set. Under the terms of the challenge, participants should exclude the corresponding training set when evaluating on one of the 75-frame sequences.

2.2 Network architectures

In this work we evaluate 4 different deep architectures for segmentation: U-Net [9,18], 2 modifications of TernausNet [11], and a modification of LinkNet [4].

In general, a U-Net-like architecture consists of a contracting path to capture context and of a symmetrically expanding path that enables precise localization (for example, see Fig. 2). The contracting path follows the typical architecture of a convolutional network with alternating convolution and pooling operations and progressively downsamples feature maps, increasing the number of feature maps per layer at the same time. Every step in the expansive path consists of an upsampling of the feature map followed by a convolution. Hence, the expansive branch increases the resolution of the output. In order to localize, upsampled

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

features, the expansive path combines them with high-resolution features from the contracting path via skip-connections [18]. The output of the model is a pixel-by-pixel mask that shows the class of each pixel. We use slightly modified version of the original U-Net model that previously proved itself very useful for segmentation problems with limited amounts of data, for example, see [9,10]. Our winning submission to the MICCAI 2017 Endoscopic Vision SubChallenge: Robotic Instrument Segmentation [13] was produced using this architecture.

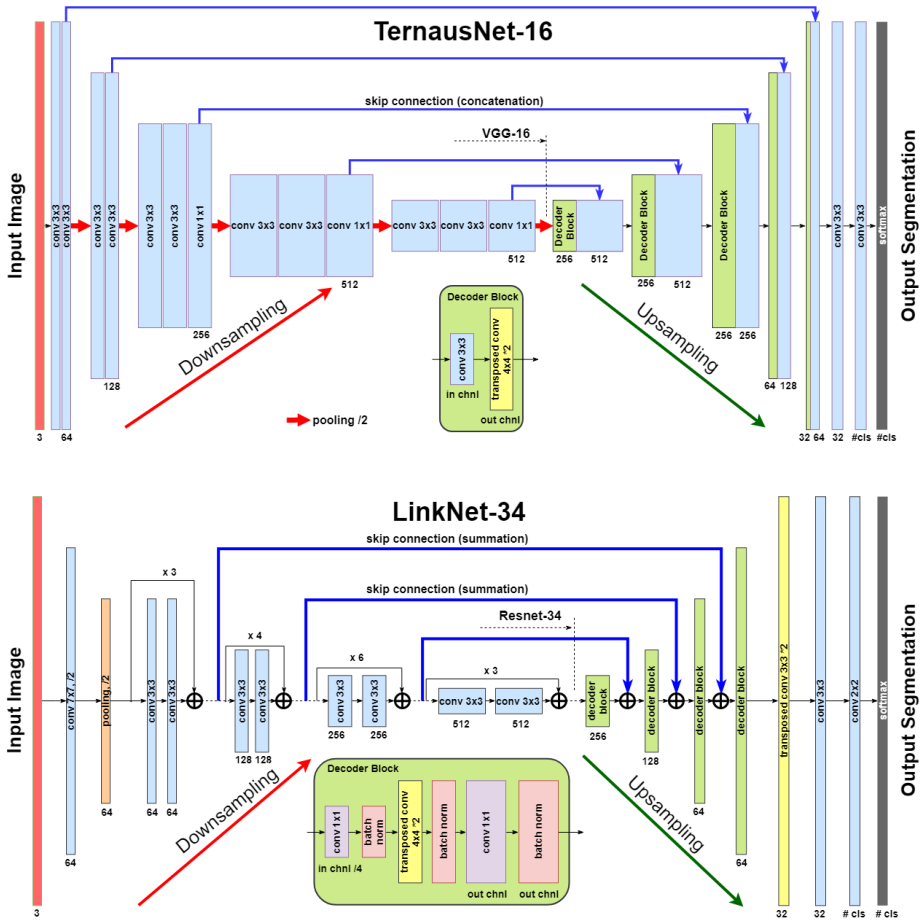

Fig. 2: These segmentation networks are based on encoder-decoder network of U-Net family. TernausNet uses pre-trained VGG16 network as an encoder, while LinkNet-34 uses pre-trained ResNet34. Each box corresponds to a multi-channel feature map. The number of channels is pointed below the box. The height of the box represents a feature map resolution. The blue arrows denote skip-connections where information is transmitted from the encoder to the decoder.

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

As an improvement over U-Net, we use similar networks with pre-trained encoders. TernausNet [11] is a U-Net-like architecture that uses relatively simple pre-trained VGG11 or VGG16 [19] networks as an encoder (see Fig. 2). VGG11 consists of seven convolutional layers, each followed by a ReLU activation function, and five max polling operations, each reducing feature map by 2. All convolutional layers have 3×3 kernels. TernausNet16 has a similar structure and uses VGG16 network as an encoder (see Fig. 2).

In contrast, LinkNet [4] model uses an encoder based on a ResNet-type architecture [8]. In this work, we use pre-trained ResNet34, see Fig. 2. The encoder starts with the initial block that performs convolution with a kernel of size 7×7 and stride 2. This block is followed by max-pooling with stride 2. The later portion of the network consists of repetitive residual blocks. In every residual block, the first convolution operation is implemented with stride 2 to provide downsampling, while the rest convolution operations use stride 1. In addition, the decoder of the network consists of several decoder blocks that are connected with the corresponding encoder block. In this case, the transmitted block from the encoder is added to the corresponding decoder block. Each decoder block includes 1×1 convolution operation that reduces the number of filters by 4, followed by batch normalization and transposed convolution to upsample the feature map.

2.3 Training

We use Jaccard index (Intersection Over Union) as the evaluation metric. It can be interpreted as a similarity measure between a finite number of sets. For two sets A and B, it can be defined as following:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

Since an image consists of pixels, the last expression can be adapted for discrete objects in the following way:

$$J = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i \hat{y}_i}{y_i + \hat{y}_i - y_i \hat{y}_i} \right) \tag{2}$$

where y_i and $\hat{y}_i$ are a binary value (label) and a predicted probability for the pixel i, correspondingly.

Since image segmentation task can also be considered as a pixel classification problem, we additionally use common classification loss functions, denoted as H. For a binary segmentation problem H is a binary cross entropy, while for a multi-class segmentation problem H is a categorical cross entropy.

The final expression for the generalized loss function is obtained by combining (2) and H as following:

$$L = H - \log J \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

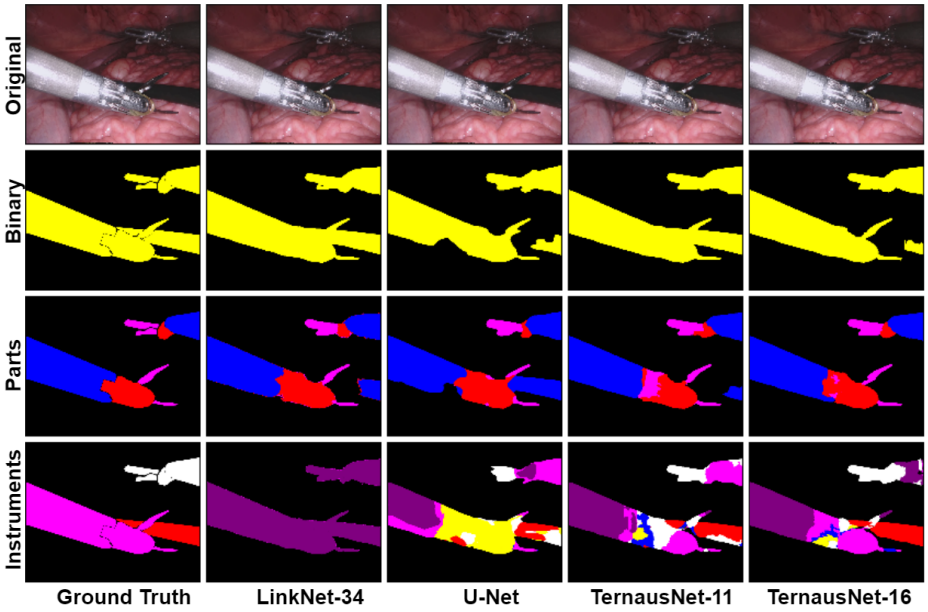

Fig. 3: Qualitative comparison between several neural network architectures implemented for a binary and multi-class segmentation.

By minimizing this loss function, we simultaneously maximize probabilities for right pixels to be predicted and maximize the intersection J between masks and corresponding predictions. We refer reader to [9] for further details.

As an output of a model, we obtain an image, in which each pixel value corresponds to a probability of belonging to the area of interest or a class. The size of the output image matches the input image size. For binary segmentation, we use 0.3 as a threshold value (chosen using validation dataset) to binarize pixel probabilities. All pixel values below the specified threshold are set to 0, while all values above the threshold are set to 255 to produce final prediction mask. For multi-class segmentation we use similar procedure, but we set different integer numbers for each class as was noted above.

3 Results

The qualitative comparison of our models both for a binary and multi-class segmentation is presented in Fig. 3 and Table 1. For the binary segmentation task the best results is achieved by TernausNet-16 providing $IoU = 0.836$ and $Dice = 0.901$. These values are the best reported in the literature up to now [7,15]. Next, we consider multi-class segmentation of different parts of instruments. As before, the best results reveals TernausNet-16 providing $IoU = 0.655$ and $Dice = 0.760$. For the multi-class class instrument segmentation task the results look less optimistic. In this case the best model is TernausNet-11 that achieves

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1: Segmentation results per task. Intersection over Union (IoU) and Dice coefficient (Dice) are in % and inference time (Time) is in ms.

	Binary segmentation			Parts segmentation			Instrument segmentation		
Model	IOU	Dice	Time	IOU	Dice	Time	IOU	Dice	Time
U-Net	75.44	84.37	93	48.41	60.75	106	15.80	23.59	**122**
TernausNet-11	81.14	88.07	142	62.23	74.25	157	**34.61**	**45.86**	173
TernausNet-16	**83.60**	**90.01**	184	**65.50**	**75.97**	202	33.78	44.95	275
LinkNet-34	82.36	88.87	**88**	34.55	41.26	**97**	22.47	24.71	177

$IoU = 0.346$ and $Dice = 0.459$ for segmentation on 7 classes. Lower performance can be explained by the relatively small dataset size. There are 7 classes and several classes appear just few times in the training dataset. Despite that we showed the best performance in the competition in this sub-category too. We are confident that this results can be drastically improved by increasing the dataset size for the corresponding problem.

When compared by the inference time, LinkNet-34 is the fastest model due to the light encoder. In the case of a binary segmentation task this network takes around 90 ms for 1280×1024 pixel image and more than twice as fast as TernausNet. The inference time was measured using one NVIDIA GTX 1080Ti GPU. A detailed comparison for the binary and multi-class tasks can be found in our GitHub repository at `https://github.com/ternaus/robot-surgery-segmentation`.

4 Conclusions

In this paper, we describe our winning solution for MICCAI 2017 Endoscopic Vision SubChallenge: Robotic Instrument Segmentation and demonstrate further improvement over that result. Our approach is originally based on U-Net network architecture that we improved using state-of-the-art semantic segmentation neural networks known as LinkNet and TernausNet. Our results shows superior performance for a binary as well as for multi-class robotic instrument segmentation. All of these networks make up end-to-end pipline and work pretty fast even for whole image resolution. We believe that our methods can lay a good foundation for similar problems of real-time surgical instrument position detection. This, in turns, can be used for the tracking and pose estimation in the vicinity of surgical scenes.

Acknowledgments Authors thank Evgeny Nizhibitsky, Alexander Buslaev, and the rest of Open Data Science community (`ods.ai`) for useful suggestions and other help aiming to the development of this work.

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Attia, M., Hossny, M., Nahavandi, S., Asadi, H.: Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3373–3378. IEEE (2017)
2. Bouget, D., Benenson, R., Omran, M., Riffaud, L., Schiele, B., Jannin, P.: Detecting surgical tools by modelling local appearance and global shape. IEEE transactions on medical imaging 34(12), 2603–2617 (2015)
3. Burgner-Kahrs, J., Rucker, D.C., Choset, H.: Continuum robots for medical applications: A survey. IEEE Transactions on Robotics 31(6), 1261–1280 (2015)
4. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. arXiv preprint arXiv:1707.03718 (2017)
5. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Xie, W., Rosen, G.L., et al.: Opportunities and obstacles for deep learning in biology and medicine. bioRxiv p. 142760 (2017)
6. Doignon, C., Nageotte, F., De Mathelin, M.: Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision. In: Dynamical Vision. pp. 314–327. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
7. García-Peraza-Herrera, L.C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E.B.V., Stoyanov, D., Vercauteren, T., Ourselin, S.: Toolnet: Holistically-nested real-time segmentation of robotic surgical tools. In: Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5717–5722. IEEE (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Iglovikov, V., Mushinskiy, S., Osin, V.: Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. arXiv preprint arXiv:1706.06169 (2017)
10. Iglovikov, V., Rakhlin, A., Kalinin, A., Shvets, A.: Pediatric bone age assessment using deep convolutional neural networks. arXiv preprint arXiv:1712.05053 (2017)
11. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
12. Kalinin, A.A., Higgins, G.A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I.D., Najarian, K., Athey, B.D.: Deep learning in pharmacogenomics: From gene regulation to patient stratification. arXiv preprint arXiv:1801.08570 (2018)
13. MICCAI 2017 Endoscopic Vision Challenge: Robotic Instrument Segmentation Sub-Challenge. `https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/`
14. Münzer, B., Schoeffmann, K., Böszörmenyi, L.: Content-based processing and analysis of endoscopic images and videos: A survey. Multimedia Tools and Applications 77(1), 1323–1362 (2018)
15. Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., Navab, N.: Deep residual learning for instrument segmentation in robotic surgery. arXiv preprint arXiv:1703.08580 (2017)
16. Pezzementi, Z., Voros, S., Hager, G.D.: Articulated object tracking by rendering consistent appearance parts. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. pp. 3940–3947. IEEE (2009)
17. Rakhlin, A., Shvets, A., Iglovikov, V., Kalinin, A.A.: Deep convolutional neural networks for breast cancer histology image analysis. arXiv preprint arXiv:1802.00752 (2018)

bioRxiv preprint doi: https://doi.org/10.1101/275867; this version posted March 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Speidel, S., Delles, M., Gutt, C., Dillmann, R.: Tracking of instruments in minimally invasive surgery for surgical skill analysis. In: Medical Imaging and Augmented Reality. pp. 148–155. Springer Berlin Heidelberg (2006)
21. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S.: Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Scientific Reports 8, 1727 (2018)