

Centerline-Diameters Data Structure for Interactive Segmentation of Tube-shaped Objects

Ilyas Sirazitdinov¹[0000–0002–5977–1534] and
Dmitry V. Dylov^{1,2,✉}[0000–0003–2251–3221]

¹ Skolkovo Institute of Science and Technology

² Artificial Intelligence Research Institute

{ilyas.sirazitdinov, d.dylov}@skoltech.ru

Abstract. Interactive segmentation techniques are in high demand in medical imaging, where the user-machine interactions are to address the imperfections of a model and to speed up the manual annotation. All recently proposed interactive approaches have kept the segmentation mask at the core, an inefficient trait if complex elongated shapes, such as wires, catheters, or veins, need to be segmented. Herein, we propose a new data structure and the corresponding click encoding scheme for the interactive segmentation of such elongated objects, without the masks. Our data structure is based on the set of centerline and diameters, providing a good trade-off between the filament-free contouring and the pixel-wise accuracy of the prediction. Given a simple, intuitive, and interpretable setup, the new data structure can be readily integrated into existing interactive segmentation frameworks.

Keywords: Interactive segmentation · Tube-shaped · Catheter segmentation · Vessel segmentation

1 Introduction

Segmentation of tube-shaped objects is an important task in medical imaging with a plethora of applications, ranging from the detection of elongated anatomical objects (blood vessels, veins, arteries, nerves) to the registration of medical devices (tubes, catheters, wires, electrodes). Co-localization of these objects may help in early detection of malpositioned invasive devices, such as central venous catheters, endotracheal tubes, and nasogastric tubes [15, 10, 19, 3, 13]. Coronary arteries disease is another notable application, where a precise segmentation of the arteries may help to identify the narrowing of the vessels [8]. In our work, we propose an interactive segmentation framework for the tube-shaped objects with the two potential use scenarios. First, the radiologists could improve their clinical routine with the tube-shaped objects by interactively correcting the imperfections with their clicks. Second, our model could be used to accelerate manual annotation of the tube-shaped objects during the data collection and labelling.

Interactive segmentation is a framework that makes it possible to control segmentation results by considering an input from a user [9, 12]. Such interactions

with the user help to localize the target objects and to adjust the predictions. The interactions can be encoded as coarse bounding boxes, text prompts, scribbles, mouse clicks, *etc.* Following the success of the click-based interactive methods in the natural domain [12], we aspired to adapt them to the medical segmentation of the elongated objects, one area of medical imaging where conventional binary segmentation can be a sub-optimal structure resulting in inefficient and inaccurate predictions [11].

Related Work. The noteworthy advances in 2D click-based methods mostly happened in the natural images domain. An iterative training and the mask guidance approach RITM was proposed in [12], where the inference-time optimization techniques were substituted by adding the previous segmentation outputs of the model during the training. An additional target crop and mask update in the localized regions was proposed in FocalClick [1]. The authors of SimpleClick [6] offered a visual transformer backbone and a symmetric patch embedding layer to encode the clicks into the backbone. A more generalized approach that can use the click interactions was proposed in Segment Anything (SAM) [4], where a foundation model, trained on a large amount of diverse data, was combined with the different interaction encodings.

Interactive segmentation is particularly important in medical imaging due to the effort required for attaining the annotation in this domain [21, 16]. A combination of a non-iterative and an interactive segmentation methods was proposed in DeepEdit [2], improving the 3D segmentation of prostatic lesions in abdominal CT. Similarly to SAM, the authors of MedSAM [7] proposed a foundation model for medical image segmentation and evaluated it using various 2D medical segmentation benchmarks. Despite the remarkable improvements in the benchmarks, both methods can not be directly applied to our task, because DeepEdit was developed for 3D, and MedSAM relied on the bounding box interactions instead of clicks.

2 Method

Our model is inspired by RITM [12] and the centerline data structure [11]. The proposed method is illustrated in Fig. 1. We use convolutional HRNet [20] model as encoder, which takes an input image of shape $(3, h, w)$, where h and w are the height and the width of the image. In contrast to RITM that encodes to input only positive and negative clicks, our encoded clicks are of shape $(4, h, w)$, where the first two channels encode *Tip 1* and *Tip 2* clicks correspondingly, and the last two channels are for the positive and the negative clicks. As in RITM, the model takes predictions from the previous click as the input. In our case, predicted mask, landmarks, and distance transform are appended to the input image and the encoded clicks. Unlike RITM, that predicts only a binary mask of the object, our model outputs the heatmap of the centerline coordinates of the shape (n_{ctr}, h, w) , where n_{ctr} is a selected number of points used to represent the centerline. In contrast to [11], that operates only with a centerline, our model outputs a mask distance transform of shape $(1, h, w)$ to fix the static width issue.

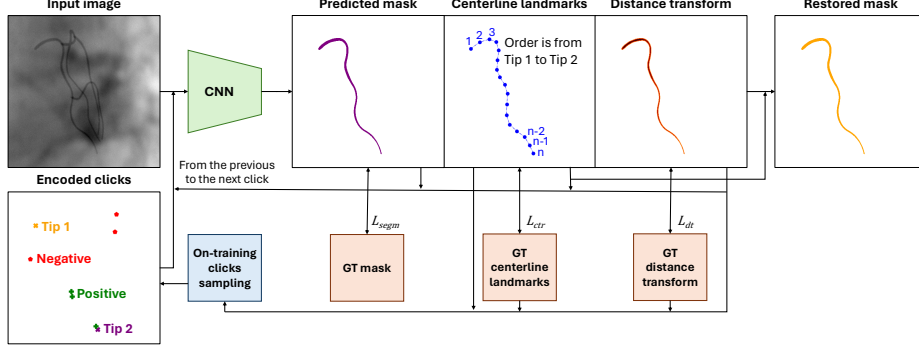


Fig. 1. Interactive segmentation scheme. CNN encoder takes input image, encoded clicks, and output from the previous click prediction (if available) and produces centerline landmarks and distance transform of the target object. Then, they are used to restore the mask of the object. Predictions, GT landmarks, and distance transform are used to generate the next positive and negative clicks during training.

In addition, we predict a mask of shape $(1, h, w)$, used only for an auxiliary loss. Both the centerline coordinates and the distance transform can be used to restore a binary mask of the target object. The ground-truth (GT) landmarks, GT distance transform, and GT mask are used to calculate the loss function and to update the encoder weights. In addition, GT data, predicted landmarks and distance transform are used for clicks sampling.

Clicks sampling. During the training, we use two click sampling schemes to simulate user interactions: initial clicks sampling [18] and on-training clicks sampling [18]. *Initial clicks sampling* corresponds to the first feeding of the model at time point $t = 0$. Similarly to minimal path methods [5], we use the tip clicks to guide the model. We randomly choose the order of the tip clicks (*Tip 1*, *Tip 2*) or (*Tip 2*, *Tip 1*) that defines the order of centerline landmarks: the first landmark corresponds to the first tip in the encoding, and the n -th point corresponds to the second tip. Then, we sample $n_{pos} \in [0, n_{max})$ positive clicks from the target foreground mask and sample $n_{neg} \in [0, n_{max})$ negative clicks from either the foreground masks of different objects or the background. Given the input image and the encoded clicks, our model outputs mask $\hat{y}_m$, centerline landmarks $\hat{y}_{ctr}$, and distance transform $\hat{y}_{dt}$. During the iterative updates at time points $t = [1; t_{max}]$ we use predictions from the previous step to generate exactly one new click according to the *on-training clicks sampling* scheme (Algorithm 1), where $\hat{y}_{coord}$ and y_{coord} are the predicted and the GT centerline coordinates, $r(c, k)$ is the function that uniformly resamples coordinates c to the fixed number of points k , $d(a, b)$ is the directed Hausdorff distance (HD) function that returns the distance between a and b and the corresponding coordinate from a . Specifically, we select the regions on a centerline with the largest HD and assign positive clicks to the false negative regions and negative clicks to the false positive regions. In case predicted centerline and GT centerline are close enough to each other

($d_{fp} > \min_{fp}$ and $d_{fn} > \min_{fn}$), we create a negative click on the foreground area with the largest error in the distance transform prediction. That click takes into account potential negative correction of the width.

Algorithm 1 On-training sampling

Input: $\hat{y}_{coord}, y_{coord}, \hat{y}_{dt}, y_{dt}, y_m$. $\triangleright$ Cent. coords, dist. transforms, GT mask
 $\hat{y}_{coord} \leftarrow r(\hat{y}_{coord}, k), y_{coord} \leftarrow r(y_{coord}, k)$
 $d_{fp}, p_{fp} \leftarrow d(\hat{y}_{coord}, y_{coord})$
 $d_{fn}, p_{fn} \leftarrow d(y_{coord}, \hat{y}_{coord})$
if $d_{fp} > d_{fn}$ and $d_{fp} > \min_{fp}$ and p_{fp} not in y_m **then**
 return: p_{fp} $\triangleright$ Negative click on a false positive region
else if $d_{fn} > \min_{fn}$ **then**
 return: p_{fn} $\triangleright$ Positive click on a false negative region
else
 $p_{dt} \leftarrow \arg \max[(y_{dt} - \hat{y}_{dt})^2 \text{ not in } y_m]$ $\triangleright$ Negative click on a foreground
 return: p_{dt} $\triangleright$ with the largest distance transform error
end if

Loss function. The loss function is an unweighted sum of three components: L_{ctr} , L_{dt} , and L_{segm} . L_{ctr} is a mean cross-entropy, calculated for each centerline point and averaged over a number of points, where only 1 pixel of GT point location has a positive label. L_{dt} is a mean squared error between the predicted distance transform of the mask and the GT. Following [12], L_{segm} is a segmentation loss (focal plus binary cross-entropy) calculated for the predicted mask. In our case, it is used only as auxiliary loss to stabilize the training. Predicted masks are not used in the output. *Inference and mask restoration.* The inference starts when the user selects a target tube-shaped object in an input image and clicks on *Tip 1* and *Tip 2*. These clicks are encoded and stacked together with the input image and are passed to the model. The model outputs centerline coordinates, the diameter of the object on the centerline coordinates, and a restored mask. Given the output, the user can add a positive click to missed regions or a negative click to false positive regions. After the encoding, that click is appended to the initial tip clicks, input image, and the current output of the model and are used as the next input for the model. The procedure is repeated until the user stops adding clicks. The centerline coordinates and the distance transform after the last click are used for the mask restoration. The mask restoration works as drawing with a spherical brush of a variable size. Namely, the brush trajectory is sampled along the predicted centerline using a linear interpolation. The brush size is sampled from the predicted distance transform as the maximum value at the ϵ -px neighborhood of the point on the centerline. The resulted trajectory of the variable brush size is the restored mask.

3 Experiments

3.1 Data

Chest x-ray catheters and tubes. We used 9085 unique images from CliP data [15, 17], with the total number of 11629 unique central venous catheters (CVC), 2994 endotracheal tubes (ETT), and 3219 nasogastric tubes (NGT).

Coronary angiography. Semi-synthetic coronary angiography was used to evaluate blood vessels segmentation. Following the methodology in [11], we generated 10000 synthetic coronary trees with a realistic cardiac background. For each image, we generated left anterior descending artery (LAD), left circumflex artery (LCX), diagonal 1 (D1), and left marginal arteries (M1). During the training, there were no distinctions between the objects from the different classes, meaning there was a single model for the chest x-rays and a single model for the coronary angiography.

Table 1. Number of clicks (NoC) required to reach certain metric value * (NoC*). Notice that our method is *on par* with the segmentation models, outperforming them in terms of object integrity (# ccs. and the Hausdorff distance).

	Chest x-ray							
	Dice		Soft Dice		Hausdorff		# ccs.	
	NoC85	NoC90	NoC90	NoC95	NoC4	NoC3	NoC2	NoC1
RITM	7.448	9.200	3.397	6.618	8.232	8.972	3.294	4.657
FocalClick	9.438	9.937	6.763	9.174	9.698	9.876	3.411	4.833
SimpleClick	8.579	9.644	4.038	8.195	8.767	9.381	2.163	3.366
Ours	8.685	8.990	5.737	8.280	4.622	6.695	2.000	2.000
	Angiography							
	Dice		Soft Dice		Hausdorff		# ccs.	
	NoC85	NoC90	NoC90	NoC95	NoC4	NoC3	NoC2	NoC1
RITM	1.542	2.505	1.198	1.937	5.069	5.260	4.513	5.433
FocalClick	3.303	5.308	1.996	4.262	7.315	7.614	3.876	5.500
SimpleClick	2.061	5.432	1.167	1.538	2.756	3.276	1.716	2.355
Ours	6.875	9.917	2.443	2.878	2.753	3.644	2.000	2.000

3.2 Model configuration

HRNet-18s [14] was used as the encoder network, taking images of size 512×512 and producing $n_{ctr} = 33$ (see Supplementary material) centerline points heatmaps, 1 distance transform map, and 1 segmentation map. All the outputs were of the same size 512×512 . The tip order was randomly sampled from two possible combinations, and the number of positive n_{pos} and negative n_{neg} clicks was randomly sampled given the normalized probability values $p_{cl+1} = p_{cl} \cdot \gamma^{cl} / \sum_{i=0}^{n_{max}} p_i$, $p_0 = 1$, with the $\gamma = 0.7$. The maximum number of clicks $n_{max} = 10$ for each click class. For the on-training sampling, we fixed $min_{pos} = 3$, $min_{neg} = 7$, and

$k = 127$. During the mask restoration we sampled the maximum diameters from the $\epsilon = 3$ neighbourhood. The coordinate extraction was an expected value calculation from the predicted logits after the softmax activation.

3.3 Results

We report the conventional number of clicks (NoC) metric that counts the average NoC needed to reach a specified segmentation metric value * (NoC*). As the prediction metrics we compute Dice, Soft Dice, undirected HD, and a number of connected components ($\# ccs$). A Soft Dice score is an extension of Dice aimed to address the impact of imperfections of the GT masks on borders caused by labelling noise and the other factors. It is calculated as a standard Dice but with the redefined confusion matrix. Given that $f(\cdot)$ is a morphological dilation, $TP = TP[f(\hat{y}), y] + TP[\hat{y}, f(y)] - TP[\hat{y}, y]$, $FP = \hat{y} \setminus f(y)$, $FN = y \setminus f(\hat{y})$. The dilation radius was set to 1 px. We compare our model with the state-of-the-art interactive segmentation frameworks: RITM [12], FocalClick[1], and SimpleClick[6] (see Supplementary material to check the hyperparameters). For a fair comparison, the same click generation strategy as in [12] was applied to all the models: the next negative or positive click is generated at the point farthest from the boundaries of the corresponding error region.

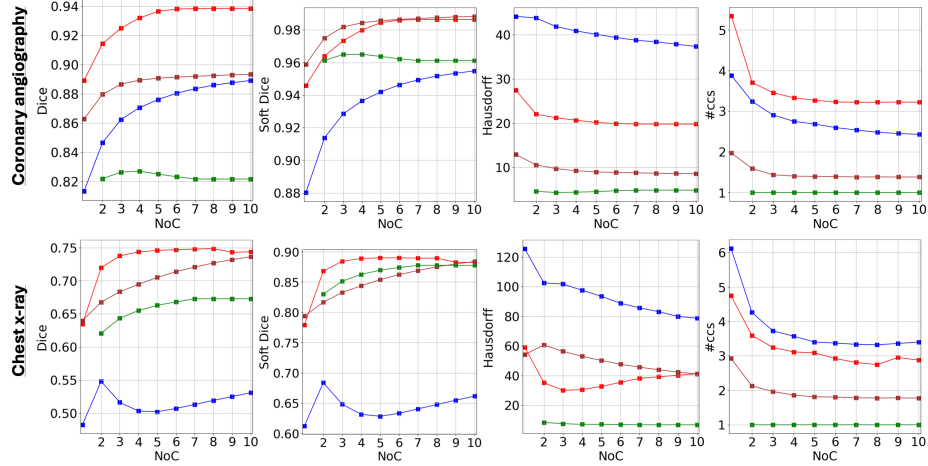


Fig. 2. Average metric values with respect to the number of clicks. *Red* - RITM, *blue* - FocalClick, *brown* - SimpleClick, *green* - ours. Notice how our method predicts true number of components early on, while keeping the segmentation scores sufficiently high.

Table 1 shows the average number of clicks each model needed to reach a certain metric value. We can notice a natural trade-off in our model performance. Our model may require more clicks to reach the high value of pixel-to-pixel segmentation metrics, such as Dice and Soft Dice, due to the point sampling limitations

and the imperfections in the restored mask. At the same time, a smaller number of clicks is required to achieve consistent segmentation (lower HD values). Moreover, in contrast to the other methods, there is an explicit guarantee that only one connected component is produced, so there is no need for additional clicks to remove false positive regions. Figure 2 shows the average segmentation metric values, given the specified number of clicks. Figs. 3 and 4 showcase the performance of the model on the images.

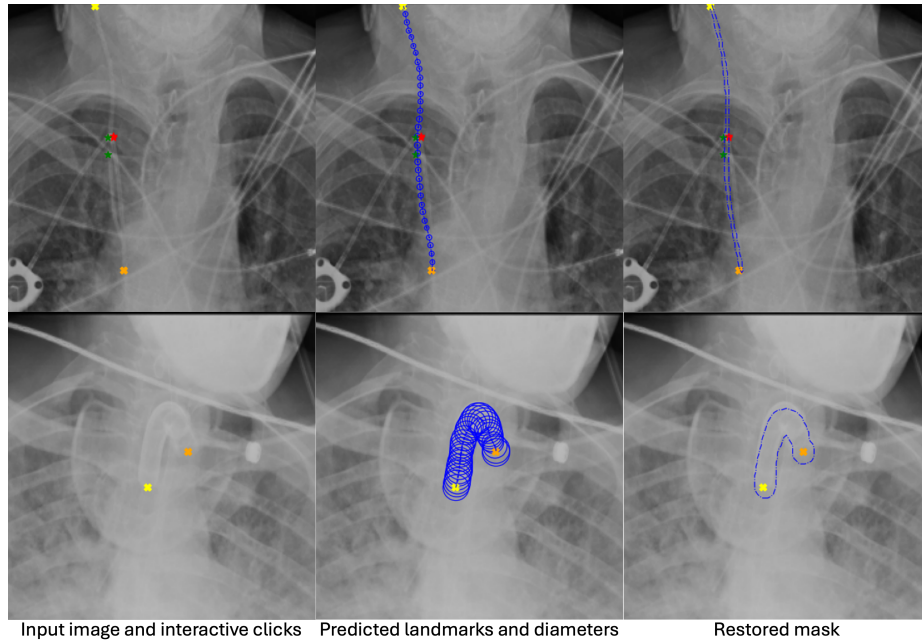


Fig. 3. Chest x-ray results. Left columns: click interactions, *yellow* for Tip 1, *orange* for Tip 2, *green* is for positive clicks, *red* is for negative clicks. Middle column: predicted centerline and sampled diameters of the target object. Right column: the restored mask.

4 Discussion

As can be noticed in Fig. 2, our model’s Dice and Soft Dice scores are not the worst, yet not the best among the compared models. They quickly plateau and fluctuate around some value. That plateau could be caused by the imperfections in the mask restoration procedure due to the sampling procedure and the potential prediction errors, another possible explanation is the noise in the labels, as the Soft Dice metric clearly suggests. At the same time, our model has already produced consistent results without any outliers after just two clicks (lower HD

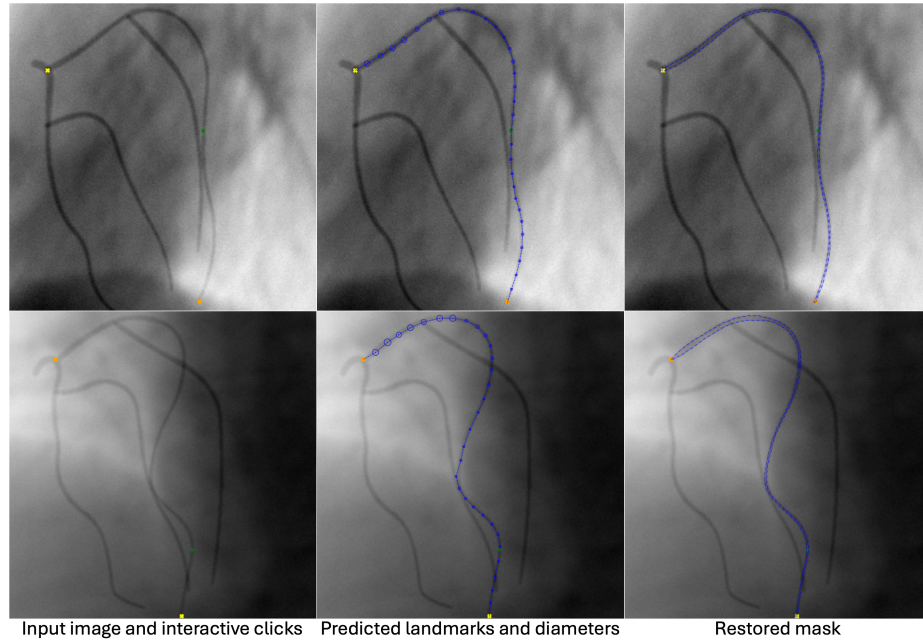


Fig. 4. Coronary angiography predictions. Left column shows click interactions: *yellow* and *orange* are for Tip 1 and Tip 2 clicks, *green* is for positive clicks, *red* is for negative clicks. Middle column show predicted centerline and sampled diameters of the target object. Right column shows restored mask

values, and always a single connected component). Notably, relatively high HD values of the masked-based methods originate from the large number of false positive connected components, which could be addressed by designing a proper connected component analysis algorithm. In addition, initial tip clicks, rather than the center-of-mass clicks, may help to better guide the model, potentially further improving the outcome of the mask-based methods. Figure 3 shows the model output and the interactive clicks. As seen from the top row, our model can resolve very challenging cases of intersecting and overlapping catheter trajectories. Namely, two positive clicks (*green*) and one negative click (*red*) helped to point out a true centerline trajectory and dismiss the path of an adjacent catheter. The bottom row of Fig. 3 shows the predictions for the tracheostomy tube. Our model managed to work with high curvature of the tube exceptionally well and produced a consistent result with minor border imperfections after the two initial clicks. Such border imperfections may potentially decrease the Dice score and the required NoC to reach high Dice values (Table 1, Fig. 2). In practice, such border imperfection may be easily fixed by moving a few control points of the contour line. Figure 4 shows the model output for the coronary angiography data. Our model successfully handled the relatively high length, high curvature, intersection regions, and overlapping regions. Remarkably, our

model required only a few clicks to achieve that result. It is important to notice, that the centerline-diameter data structure for segmentation is not limited to RITM, but can be readily integrated into different interactive segmentation frameworks with minor changes in the click encoding scheme and the loss function calculation, if needed. Moreover, the centerline as a sequential structure can be predicted in an autoregressive way, potentially further improving the model’s accuracy. It is also not limited to the click-based approaches and could be combined with more suitable interactions for the tube objects, *e.g.* rough scribbles or approximate outlines of the objects.

Conclusions. We addressed the problem of interactive segmentation of elongated tubular objects, validating our solution both on anatomical objects and on invasive medical devices. We combined the most recent techniques from the interactive models in the natural domain [12] and enhanced the centerline-based data structure [11] to be more suitable for tubular objects. Moreover, our model fixed the static width issue of [11] by explicit regression of the distance transform and the diameter sampling. Our data structure offers a trade-off between the object integrity and the pixel-wise accuracy of segmentation; thus, it is an ideal candidate for the integration with various existing interactive medical frameworks.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1300–1309 (2022)
2. Diaz-Pinto, A., Mehta, P., Alle, S., Asad, M., Brown, R., Nath, V., Ihsani, A., Antonelli, M., Palkovics, D., Pinter, C., et al.: Deepedit: deep editable learning for interactive segmentation of 3d medical images. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 11–21. Springer (2022)
3. Frid-Adar, M., Amer, R., Greenspan, H.: Endotracheal tube detection and segmentation in chest radiographs using synthetic data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 784–792. Springer (2019)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
5. Liao, W., Wörz, S., Kang, C.K., Cho, Z.H., Rohr, K.: Progressive minimal path method for segmentation of 2d and 3d line structures. *IEEE transactions on pattern analysis and machine intelligence* **40**(3), 696–709 (2017)
6. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22290–22300 (2023)

7. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications* **15**(1), 654 (2024)
8. Pan, L.S., Li, C.W., Su, S.F., Tay, S.Y., Tran, Q.V., Chan, W.P.: Coronary artery segmentation under class imbalance using a u-net based architecture on computed tomography angiography images. *Scientific Reports* **11**(1), 1–7 (2021)
9. Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., Erickson, B.J.: Interactive segmentation of medical images through fully convolutional neural networks. *arXiv preprint arXiv:1903.08205* (2019)
10. Sirazitdinov, I., Lenga, M., Baltruschat, I.M., Dylov, D.V., Saalbach, A.: Landmark constellation models for central venous catheter malposition detection. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. pp. 1132–1136. IEEE (2021)
11. Sirazitdinov, I., Saalbach, A., Schulz, H., Dylov, D.V.: Bi-directional encoding for explicit centerline segmentation by fully-convolutional networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 693–703. Springer (2022)
12. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: *2022 IEEE International Conference on Image Processing (ICIP)*. pp. 3141–3145 (2022). <https://doi.org/10.1109/ICIP46576.2022.9897365>
13. Subramanian, V., Wang, H., Wu, J.T., Wong, K.C., Sharma, A., Syeda-Mahmood, T.: Automated detection and type classification of central venous catheters in chest x-rays. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 522–530. Springer (2019)
14. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514* (2019)
15. Tang, J.S., Seah, J.C., Zia, A., Gajera, J., Schlegel, R.N., Wong, A.J., Gai, D., Su, S., Bose, T., Kok, M.L., et al.: Clip, catheter and line position dataset. *Scientific Data* **8**(1), 1–7 (2021)
16. Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al.: Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging* **37**(7), 1562–1573 (2018)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2097–2106 (2017)
18. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S.: Deep interactive object selection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 373–381 (2016)
19. Yi, X., Adams, S.J., Henderson, R.D., Babyn, P.: Computer-aided assessment of catheters and tubes on radiographs: How good is artificial intelligence for assessment? *Radiology: Artificial Intelligence* **2**(1), e190082 (2020)
20. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation (2020)
21. Zhao, F., Xie, X.: An overview of interactive medical image segmentation. *Annals of the BMVA* **2013**(7), 1–22 (2013)