ICIC Express Letters
Volume **16**, Number **7**, July **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **795–802**

EFFICIENT INSTANCE SEGMENTATION ANNOTATION USING SCRIBBLES AND SUPERPIXEL

Suryadiputra Liawatimena[1,2,*], Edi Abdurahman[1], Agung Trisetyarso[1]
Antoni Wibowo[2] and Josef Budi Indrasworo[2]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[2]Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ edia; josef.indrasworo }@binus.ac.id; { atrisetyarso; anwibowo }@binus.edu
*Corresponding author: suryadi@binus.edu

Received July 2021; accepted October 2021

Abstract. *In relation to the development of an object detection system based on deep learning, preparation of instance segmentation dataset takes a lot of effort, time, and costs, concerning activities to annotate object boundaries and provide adequate data. To be able to focus on a problem one tries to solve, an efficient method for dataset annotation is needed. One method which tries to address this is FreeLabel, where users create freehand scribbles to generate segmentation mask without specifying exact and detailed object boundaries. However, generating segmentation mask still needs a while and it occurs number of times as users create scribbles, get the resulting mask, and improve them accordingly by modifying the scribbles. The goal of this study was to find similar approach with reduced segmentation mask generation time, so users do not have to wait long for feedback in form of resulting mask as they shape final segmentation mask interactively. Measured against reference dataset containing medium sized images and corresponding ground truths, method used in this study reduced average generation time from about 8 seconds to under 1 second while still maintaining to have comparable quality and user effort.*

Keywords: Image annotation, Instance segmentation, Dataset preparation, Object detection, Image segmentation, Deep learning

1. **Introduction.** Computer vision tasks using deep learning consist of several tasks, one of which is instance segmentation. Other tasks include image classification, object detection (localization), semantic segmentation, image captioning, and key point detection. The instance segmentation task takes an image as input and then predicts the class and segmentation mask of certain instances (of already defined classes) in it. This provides detailed shape descriptions of target instances compared to the bounding box. Some applications of instance segmentation in diverse fields are [1-5]. Various deep learning methods related to instance-level visual recognition tasks have been explored. Mask R-CNN (Region Based Convolutional Neural Network) [6] showed top result in Microsoft COCO (Common Objects in Context) challenge for instance segmentation track and became popular method used in other researches. Researchers later tried to improve it in several papers like [7-11].

Supervised deep learning method needs a dataset that is used for model training and testing. The deep learning model needs adequate training data to be able to perform well. Additionally, more training data tends to improve model accuracy. Several image datasets that contain images along with ground truth annotations are publicly available for research. Some public datasets which have segmentation annotations are Pascal Visual

DOI: 10.24507/icicel.16.07.795

Object Classes (VOC) challenge dataset [12], SUN (Scene UNderstanding) database [13], Microsoft COCO [14], and ADE20K [15]. Microsoft COCO dataset is a collection of about 328,000 images with rich contextual relationships among objects contained in them. The annotation process of more than 2.5 million objects (of 91 categories) in the image collection was crowdsourced and done by workers at Amazon's Mechanical Turk (AMT).

Many efforts are needed to prepare instance segmentation datasets because of the abundant data to be annotated and the tedious annotation task. Some approaches have been proposed in various research fields to reduce the annotation process's burden and difficulty in general as in [16-20]. Related to instance segmentation annotation, several tools or techniques like Polygon-RNN (Recurrent Neural Network) [21], Polygon-RNN++ [22], ByLabel [23], Superlabel [24], GTCreator [25] were proposed. FreeLabel, a web interface which allows user to create segmentation mask using only a few freehand scribbles, is a semi-automated interactive segmentation tool [26] and was used to create The Fruit Flowers dataset [27]. It displayed good results when tested against selected datasets, but the process that grows scribbles into segmentation mask was still expected to be accelerated [28]. Meanwhile, it was also observed that better outcomes were obtained from user who frequently grew scribbles to interactively refine the generated mask.

This paper presents a study to improve on existing FreeLabel approach. It eases annotation process in that user does not have to specify exact object boundary, while still has flexibility to shape the resulting segmentation mask. However, in some machines, it will take some seconds to process and produce segmentation mask. Meanwhile, for one image user it will generally need several corresponding processes to be able to repeatedly refine the generated segmentation mask. This study tried to reduce user waiting time by accelerating segmentation mask generation process, while still maintaining FreeLabel's quality of result and modest effort. The rest of this work is presented as follows. In Section 2, method used in this study is described. Section 3 provides dataset description, experimental work, and discussion of results of the experiment. The conclusion of this work is given in Section 4.

2. Method.

This study's three measurements were Intersection over Union (IoU), number of scribbles, and segmentation mask generation time. IoU was used to measure the quality of the annotation process. The number of scribbles and segmentation mask generation time were measurements of annotation effort. The number of scribbles is the amount of scribbles created in an annotation process. Segmentation mask generation time, measured in seconds, is amount of time needed to convert user defined scribbles into segmentation mask.

IoU or Jaccard index (Jaccard similarity index) measures annotation accuracy by estimating the similarity between annotation results and ground truth. Given a source image and a target image, the Jaccard index can measure overlap between the images. Jaccard index (IoU) is defined as in (1). Annotation results and ground truth segmentation mask were compared to get intersection and union mask for each instance. Number of mask pixels from intersection and union mask were counted and then used to compute IoU.

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

While the overall annotation process is basically the same as in FreeLabel, the main user interface was changed a bit. Scribble is implemented as vector graphics based on VGG Image Annotator (VIA) code [29]. Table 1 shows some differences between original FreeLabel and annotation tool used in this study. Figure 1 illustrates steps of the annotation process using the tool. On the left part of the figure, polyline shaped scribbles were created, and each of them had its corresponding object defined (in this case either background, person, or fish). When user decided to generate mask, a PNG image containing

Table 1. Some UI differences between FreeLabel and its implementation in this study

Type	Original FreeLabel	Implementation in this study
Scribble form	Free scribbles or lines	Polygon outline or polyline shapes
Scribble implementation	A group of pixels	As vector graphics
Scribble modification	Using eraser scribble to erase some of the existing pixels	Delete or move the shape, or modify shape by moving its vertices

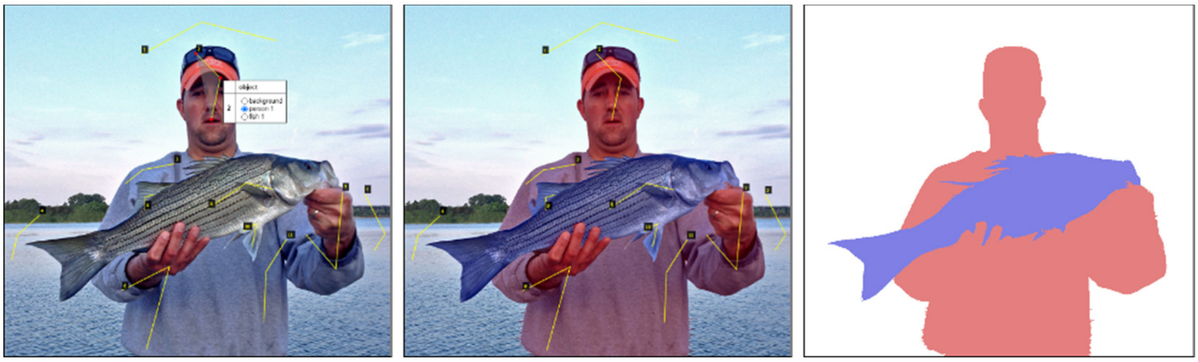

Figure 1. Example of creating scribbles and getting generated segmentation mask

segmentation mask was created (as seen on the right part of the figure) and was set as an overlay to the annotated image (middle part).

The proposed method still used SNIC (Simple Non-Iterative Clustering) to create superpixels of input image. Superpixel is collection of adjacent pixels. Each superpixel containing scribbles created by user will be labelled as the object label of the scribbles only if there is no conflict. From that initial labelling, label of other superpixels (except for superpixels having conflict) will be defined based on average color similarity with already labelled superpixel. Color similarity was defined as distance of 2 colors in CIELAB color space.

For superpixel having conflict (or superpixel in which there are more than 1 object label as defined by scribbles), each pixel in it is labelled separately. Approach to classifying these pixels was similar to the approach used in FreeLabel. For each superpixel, pixels which are part of user scribbles are labelled according to scribble's label. Those pixels then are used as seeds to determine label of other pixels in corresponding superpixel using FreeLabel approach (which in turn based on SNIC [30] algorithm).

With number of superpixels of an image significantly less than number of its pixels, the approach in this study reduced processing time because changes in scribbles did not require superpixels regeneration. However, creating segmentation based on superpixel alone mostly will not give satisfying result. Mechanism to resolve conflict in superpixel labelling was used to solve that by allowing user to shape result to get improved segmentation mask.

Implementation of FreeLabel approach in the proposed method differed slightly in that all pixels were used as seeds, contrasting with several pixels selected randomly as seeds. This also made the resulting segmentation mask fixed instead of changing a bit every time it was generated. There was only single superpixels generation to obtain the final segmentation mask result, in contrast with multiple superpixels generation that needed further processing.

Not all the processes mentioned before needed to be run every time scribbles were converted to segmentation mask. Some processes like generation of superpixels only had to be executed once for each image. Result of conflict resolution also was stored, so that if scribbles in certain superpixel did not change, there was no need for that processing.

3. **Results and Discussion.** Experiment was conducted where each image in test dataset was annotated using FreeLabel and the proposed method. Figure 2 illustrates the annotations (scribbles) and results using FreeLabel and the proposed method for one of the images in test dataset. Red line represented scribbles for foreground fish object, and black line represented scribbles for background. The laptop's hardware specifications used to carry out annotation processes were Intel i5 CPU, 8Gb RAM, and Intel HD Graphics 3000.

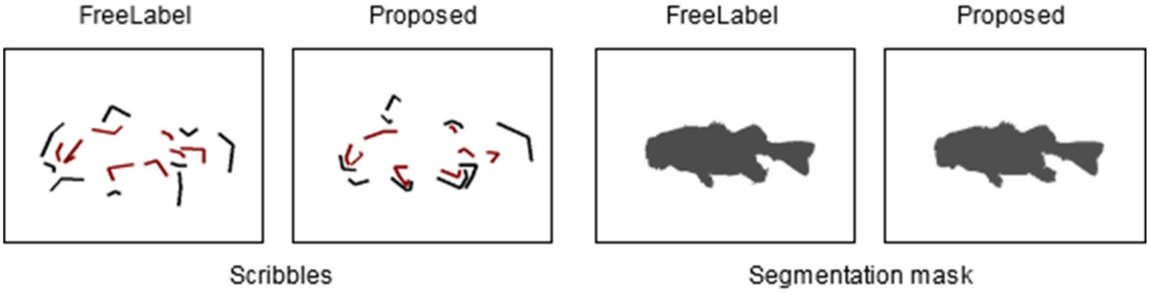

Figure 2. (color online) Illustration of scribbles and generated segmentation masks

To measure performance of annotation method, a light reference dataset was prepared. The dataset consisted of images and the corresponding segmentation mask ground truths. All images were images depicting fish catch. Example of an image ("Milford Lake Fish" by Acorns Resort, https://www.flickr.com/photos/acornsresortkansas/15581857861/, CC BY 2.0) and its ground truth from the reference dataset are given in Figure 3. Each image was medium (about 800 × 600 pixels) or smaller sized image. The reference dataset had 10 images. Each image contained only 1 fish object to be annotated.

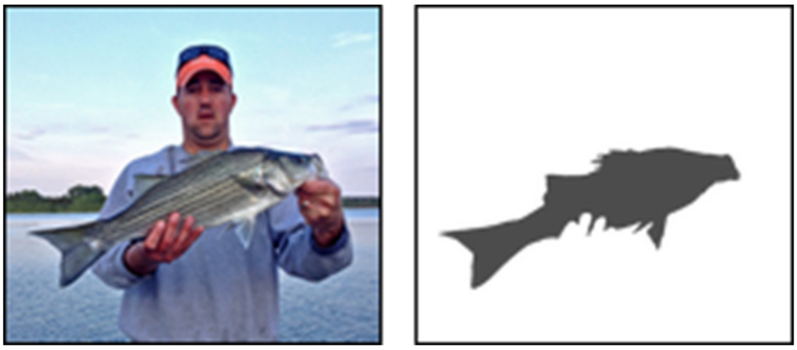

Figure 3. Example of image (left) and ground truth (right) in reference dataset

The type of experiment used in this study was one-group pretest-posttest where a group of sample data is tested before and after treatment. The treatment here was method improvement, so pretest was conducted using FreeLabel, and posttest was conducted using the proposed method. The group of sample data was images in reference dataset, and the test itself was instance segmentation annotation on each image using corresponding method. With the annotation method as the input (independent) variable, the outcome variables of experiment were the 3 measurements mentioned in Section 2. Table 2 summarizes experiment managed in this study. Expected result of the experiment was that the proposed method reduced mask generation time, but it did not significantly reduce similarity index between resulting mask and ground truth, and it did not significantly

Table 2. Experiment design

Sample data	Pretest	Posttest	Outcome
Images in reference dataset	Annotation using FreeLabel	Annotation using the proposed method	IoU, number of scribbles, segmentation mask generation time (in second)

increase the number of scribbles needed. To confirm this, these 3 assessments would be carried out:

1) check if the proposed method reduced segmentation mask generation time,

2) check if the proposed method did not significantly reduce similarity index between resulting mask and ground truth, and

3) check if the proposed method did not significantly increase number of scribbles needed.

Intersection over Union (IoU) describing likeness between result and ground truth was a measurement to control the quality of the proposed method. It could not be significantly lower than the one produced by the compared method. Number of scribbles was another measurement to control the quality of the proposed method. It could not be significantly higher than the one produced by the compared method. Figure 4 shows IoU comparison and number of scribbles comparison between 2 methods for each image annotation. Both methods produced similar results for those 2 measurements. One method gave slightly higher results on certain images while providing slightly lower results on others.

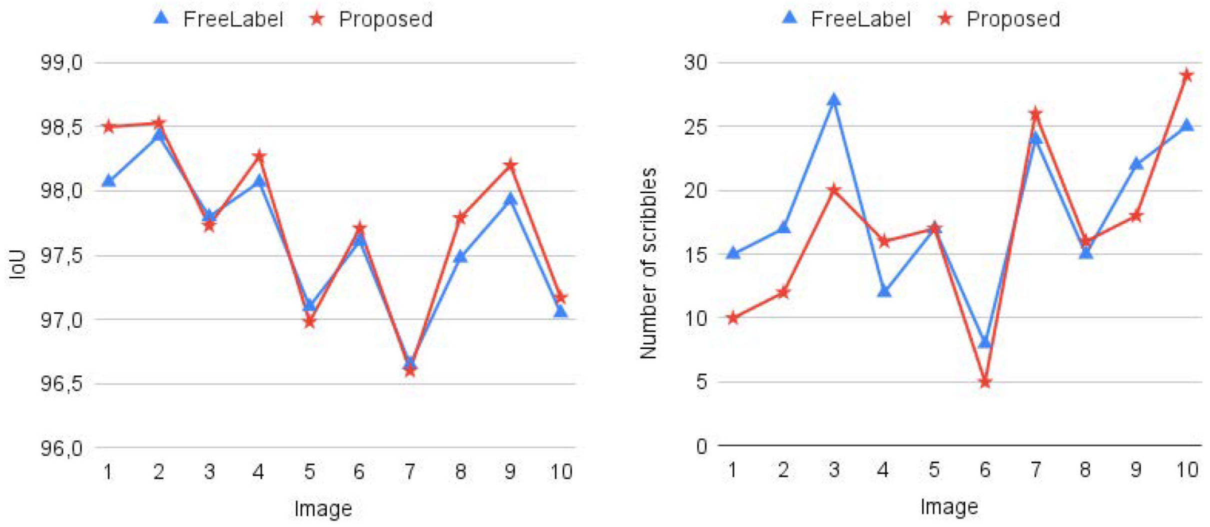

Figure 4. Per image comparison of IoU (left) and number of scribbles (right)

Summary of these measurements then can be seen in Table 3 and Table 4. Looking at average increase of 0.13% in IoU, although there was also decrease of 0.05% on minimum value, it could be said that the proposed method did not produce significantly lower similarity index. Looking at average decrease of −9.34% in number of scribbles, although there was also increase of 7.41% (2 more scribbles) on maximum value, it could be said that the proposed method did not require significantly higher number of scribbles.

Segmentation mask generation time was the measurement in concern for the proposed method. It was expected to be lower than the one performed by the compared method. Figure 5 shows generation time comparison between FreeLabel and the proposed method for each image annotation. The proposed method needed less time to generate segmentation mask. Summary of measurement then can be seen in Table 5. Looking at average decrease −93.45%, although there was additional initiation time for creating information

TABLE 3. Summary of similarity index comparison

	Intersection over Union (%)		
	FreeLabel	**Proposed**	**Changes (%)**
Average	97.62	97.75	0.13
Standard deviation	0.55	0.66	19.17
Maximum	98.43	98.53	0.10
Minimum	96.65	96.60	−0.05

TABLE 4. Summary of number of scribbles comparison

	Number of scribbles		
	FreeLabel	**Proposed**	**Changes (%)**
Average	18.20	16.05	−9.34
Standard deviation	6.12	7.01	14.49
Maximum	27.00	29.00	7.41
Minimum	8.00	5.00	−37.50

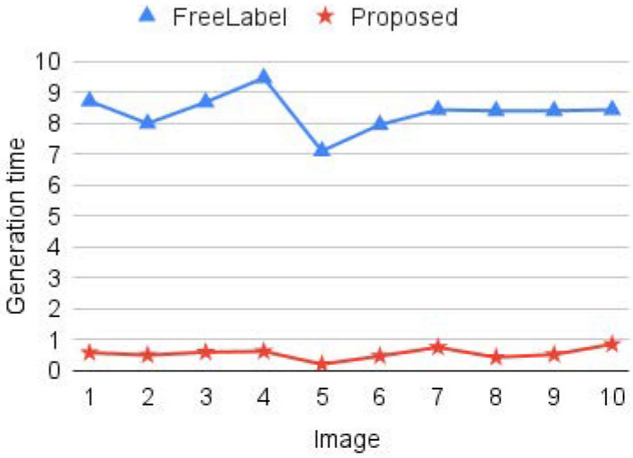

FIGURE 5. Per image comparison of segmentation mask generation time

TABLE 5. Summary of segmentation mask generation time comparison

	Mask generation time (second)			Initiation time (second)	
	FreeLabel	**Proposed**	**Changes (%)**	**FreeLabel**	**Proposed**
Average	8.36	0.55	−93.45	–	2.96
Standard deviation	9.73	0.18	−70.90	–	0.84
Maximum	9.46	0.85	−91.05	–	3.82
Minimum	7.10	0.20	−97.20	–	0.78

that will be used later (superpixel, graph of superpixels, etc.) once per image, it could be said that the proposed method performed lower segmentation mask generation time.

Based on previous evaluation of the 3 measurements, related to question of study as can be seen in Table 6, all 3 assessments were successful. They confirmed the expectation of this study. This meant that the proposed method reduced segmentation mask generation time without sacrificing much the quality of result and effort needed from the compared method.

Table 6. Result of experiment

Expectation		
The proposed method reduced mask generation time, but did not significantly reduce similarity index between resulting mask and ground truth, and did not significantly increase the number of scribbles needed		Confirmed
Assessment		
1	The proposed method reduced segmentation mask generation time	True
2	The proposed method did not significantly reduce similarity index between resulting mask and ground truth	True
3	The proposed method did not significantly increase the number of scribbles needed	True

4. **Conclusions.** Both FreeLabel and the proposed method performed well against reference dataset with result of more than 96% similarity for each image. Both methods required similar effort based on number of scribbles and produced similar results based on IoU measurement. With comparable quality and effort, the proposed approach generally needed less time to produce result in response to modified user's scribbles. Therefore, it can be useful as an alternative method that will provide users with quicker feedback for each iteration to shape final segmentation mask and that could be potentially improved further.

Even if it performed well enough, not all images could be annotated easily, and the generated segmentation mask could be improved further. For that reason, future works may include studies about how to enhance annotation process or the result. Moreover, with various researches in superpixel domain, next step could be to conduct a study to compare different superpixel algorithms in order to find the most appropriate one. Integration with or introducing other method may as well help enhance the process or result because some boundaries might be easier to be defined by certain method. For instance, improvement might be achieved by manually defining object boundaries on certain parts of the object.

REFERENCES

[1] Y. Qiao, M. Truman and S. Sukkarieh, Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming, *Comput. Electron. Agric.*, vol.165, DOI: 10.1016/j.compag.2019.104958, 2019.

[2] J. Y. Chiao, K. Y. Chen, K. Y.-K. Liao, P. H. Hsieh, G. Zhang and T. C. Huang, Detection and classification the breast tumors using Mask R-CNN on sonograms, *Med. (United States)*, vol.98, no.19, pp.1-5, 2019.

[3] Z. Yang, Y. Yuan, M. Zhang, X. Zhao, Y. Zhang and B. Tian, Safety distance identification for crane drivers based on Mask R-CNN, *Sensors (Switzerland)*, vol.19, no.12, 2019.

[4] W. Jia, Y. Tian, R. Luo, Z. Zhang, J. Lian and Y. Zheng, Detection and segmentation of overlapped fruits based on optimized Mask R-CNN application in apple harvesting robot, *Comput. Electron. Agric.*, vol.172, DOI: 10.1016/j.compag.2020.105380, 2020.

[5] C. Yu et al., Segmentation and measurement scheme for fish morphological features based on Mask R-CNN, *Inf. Process. Agric.*, vol.7, no.4, pp.523-534, 2020.

[6] K. He, G. Gkioxari, P. Dollár and R. Girshick, Mask R-CNN, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.42, no.2, pp.386-397, 2020.

[7] Z. Cai and N. Vasconcelos, Cascade R-CNN: High quality object detection and instance segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.43, no.5, pp.1483-1498, 2021.

[8] M. Wu et al., Object detection based on RGC Mask R-CNN, *IET Image Process.*, vol.14, no.8, pp.1502-1508, 2020.

[9] Z. Huang, L. Huang, Y. Gong, C. Huang and X. Wang, Mask scoring R-CNN, *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp.6402-6411, 2019.

[10] J. H. Shu, F. D. Nian, M. H. Yu and X. Li, An improved Mask R-CNN model for multiorgan segmentation, *Math. Probl. Eng.*, 2020.

[11] Y. Tian, G. Yang, Z. Wang, E. Li and Z. Liang, Instance segmentation of apple flowers using the improved Mask R-CNN model, *Biosyst. Eng.*, vol.193, pp.264-278, 2020.

[12] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, The pascal visual object classes challenge: A retrospective, *Int. J. Comput. Vis.*, vol.111, no.1, pp.98-136, 2015.

[13] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba and A. Oliva, SUN database: Exploring a large collection of scene categories, *Int. J. Comput. Vis.*, vol.119, no.1, pp.3-22, 2016.

[14] T.-Y. Lin et al., Microsoft COCO: Common objects in context, *European Conference on Computer Vision*, pp.740-755, 2014.

[15] B. Zhou et al., Semantic understanding of scenes through the ADE20K dataset, *Int. J. Comput. Vis.*, vol.127, no.3, pp.302-321, 2019.

[16] E. Lughofer, Hybrid active learning for reducing the annotation effort of operators in classification systems, *Pattern Recognit.*, vol.45, no.2, pp.884-896, 2012.

[17] Z. Yu, C. Li, J. Wu, J. Cai, M. N. Do and J. Lu, Action recognition in still images with minimum annotation efforts, *IEEE Trans. Image Process.*, vol.25, no.11, pp.5479-5490, 2016.

[18] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller and V. Ferrari, Extreme clicking for efficient object annotation, *Proc. of IEEE Int. Conf. Comput. Vis.*, pp.4940-4949, 2017.

[19] Z. Xing, M. Zang and Y. Zhang, Learning regularized multi-view structured sparse representation for image annotation, *International Journal of Innovative Computing, Information and Control*, vol.14, no.4, pp.1267-1283, 2018.

[20] D. Zhang, J. Han, Y. Zhang and D. Xu, Synthesizing supervision for learning deep saliency network without human annotation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.42, no.7, pp.1755-1769, 2019.

[21] L. Castrejón, K. Kundu, R. Urtasun and S. Fidler, Annotating object instances with a Polygon-RNN, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4485-4493, DOI: 10.1109/CVPR.2017.477, 2017.

[22] D. Acuna, H. Ling, A. Kar and S. Fidler, Efficient interactive annotation of segmentation datasets with Polygon-RNN++, *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp.859-868, 2018.

[23] X. Qin, S. He, Z. Zhang, M. Dehghan and M. Jagersand, ByLabel: A boundary based semi-automatic image annotation tool, *Proc. of 2018 IEEE Winter Conf. Appl. Comput. Vision (WACV2018)*, pp. 1804-1813, 2018.

[24] M. Wigness, *SuperLabel: A Superpixel Labeling Interface for Semantic Image Annotation*, Army Research Lab Adelphi MD Adelphi, United States, 2018.

[25] J. Bernal et al., GTCreator: A flexible annotation tool for image-based datasets, *Int. J. Comput. Assist. Radiol. Surg.*, vol.14, no.2, pp.191-201, 2019.

[26] H. R. Roth, D. Yang, Z. Xu, X. Wang and D. Xu, Going to extremes: Weakly supervised medical image segmentation, *Mach. Learn. Knowl. Extr.*, vol.3, no.2, pp.507-524, 2021.

[27] Y. Lu and S. Young, A survey of public datasets for computer vision tasks in precision agriculture, *Comput. Electron. Agric.*, vol.178, DOI: 10.1016/j.compag.2020.105760, 2020.

[28] P. A. Dias, Z. Shen, A. Tabb and H. Medeiros, FreeLabel: A publicly available annotation tool based on freehand traces, *Proc. of 2019 IEEE Winter Conf. Appl. Comput. Vision (WACV2019)*, pp.21-30, 2019.

[29] A. Dutta and A. Zisserman, The VIA annotation software for images, audio and video, *Proc. of the 27th ACM Int. Conf. Multimed. (MM2019)*, pp.2276-2279, 2019.

[30] R. Achanta and S. Süsstrunk, Superpixels and polygons using simple non-iterative clustering, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4895-4904, DOI: 10.1109/CVPR.2017.520, 2017.