RDMP: A REFERENCE DETECTION AND MASK PROPAGATION PIPELINE FOR NUMEROUS 3D GLOMERULI SEGMENTATION IN LARGE VOLUMES

Ziling Lin[1†] *Zehua Li*[2†] *Zifan Chen*[1] *Yao Lu*[2]
Fangxu Zhou[3] *Bin Dong*[4,5,6] *Li Zhang*[1,6⋆] *Haifeng Li*[4⋆]

[1] Center for Data Science, Peking University, Beijing, China
[2] Peking University First Hospital, Beijing, China
[3] College of Future Technology, Peking University, Beijing, China
[4] Beijing International Center for Mathematical Research, Peking University, Beijing, China
[5] Center for Machine Learning Research, Peking University, Beijing, China
[6] National Biomedical Imaging Center, Peking University, Beijing, China

ABSTRACT

Quantitative analysis of numerous repetitive structures in large volumes is one of the most challenging tasks in biomedical imaging, primarily due to its intensive manual requirements. Kidney glomeruli exemplify this challenge due to their dispersion across the cortical region, high density and microscopic size. To address this, we present a Reference Detection and Mask Propagation (RDMP) pipeline that combines object detection with 2D and 3D mask propagation, gradually achieving efficient segmentation of the entire volume by using a 2D reference example. Our method demonstrates robust performance, achieving a Dice Similarity Coefficient (DSC) of 0.9096 on microCT data, and when generalized to Synchrotron Radiation X-ray (SRX) data, reaching a DSC of 0.8431 with only minimal bounding box fine-tuning. Overall, this pipeline shows strong potential for efficient, large-scale kidney analysis, significantly reducing labor demands and accelerating kidney disease research.

***Index Terms*—** Large-scale Volume, Reference Detection, Mask Propagation, 3D Glomeruli Segmentation

1. INTRODUCTION

In biomedical image analysis, handling large-scale repetitive structures poses a significant challenge, primarily due to the substantial manual resources required for annotation [1]. The kidney, essential for waste filtration and fluid balance maintenance, is closely associated with various diseases when its functions are impaired. Accurate diagnosis and treatment of kidney-related diseases require the creation of a detailed kidney atlas with precise segmentation of its components [2]. A critical step is the segmentation of glomeruli, which are microscopic structures fundamental to the kidney's filtration process [3]. With approximately 1,000,000 glomeruli in each human kidney, their comprehensive segmentation is a resource-intensive task, highlighting the need for efficient, reliable methods for segmenting repetitive structures.

A straightforward approach to segmenting large repetitive structures involves employing object detection models to locate potential targets before performing segmentation. Existing methods typically fall into two categories. The first integrates segmentation modules directly into object detection networks, exemplified by models like Mask R-CNN [4], YOLO [5] and DETR [6]. The second approach, often used in interactive models such as T-Rex [7] and T-Rex2 [8], applies detected bounding boxes as prompts for interactive segmentation models such as SAM [9]. However, both approaches are vulnerable to detection errors, as segmentation accuracy depends heavily on detection results. The tight coupling of detection and segmentation poses particular challenges in accurately segmenting numerous small, repetitive structures like glomeruli, especially when initial detections are inaccurate.

To address these limitations, we propose a three-stage Reference Detection and Mask Propagation (RDMP) pipeline, illustrated in Fig. 1(a), consisting of sequential stages: 2D reference detection, 2D mask propagation, and 3D mask propagation. In the 2D detection stage, by giving a reference object, we enhance the detection model with a visual-prompt branch and a multi-scale cross-attention (MSCA) module to improve detection accuracy for small, dense targets. Following detection, we introduce 2D and 3D mask propagation based on PAM (Propagating Anything Model) [10], a propagation-based general segmentation model for 3D medical images. Specifically, in the 2D stage, reference masks are propagated to segment the detected regions. The 3D stage then extends these masks to adjacent slices along the vertical axis to complete the full 3D volume. In our pipeline, each successive stage refines its predecessor's results: false positives from

† Equal contribution.

⋆ Corresponding authors: Li Zhang(zhangli_pku@pku.edu.cn), Haifeng Li(lihf@bicmr.pku.edu.cn)

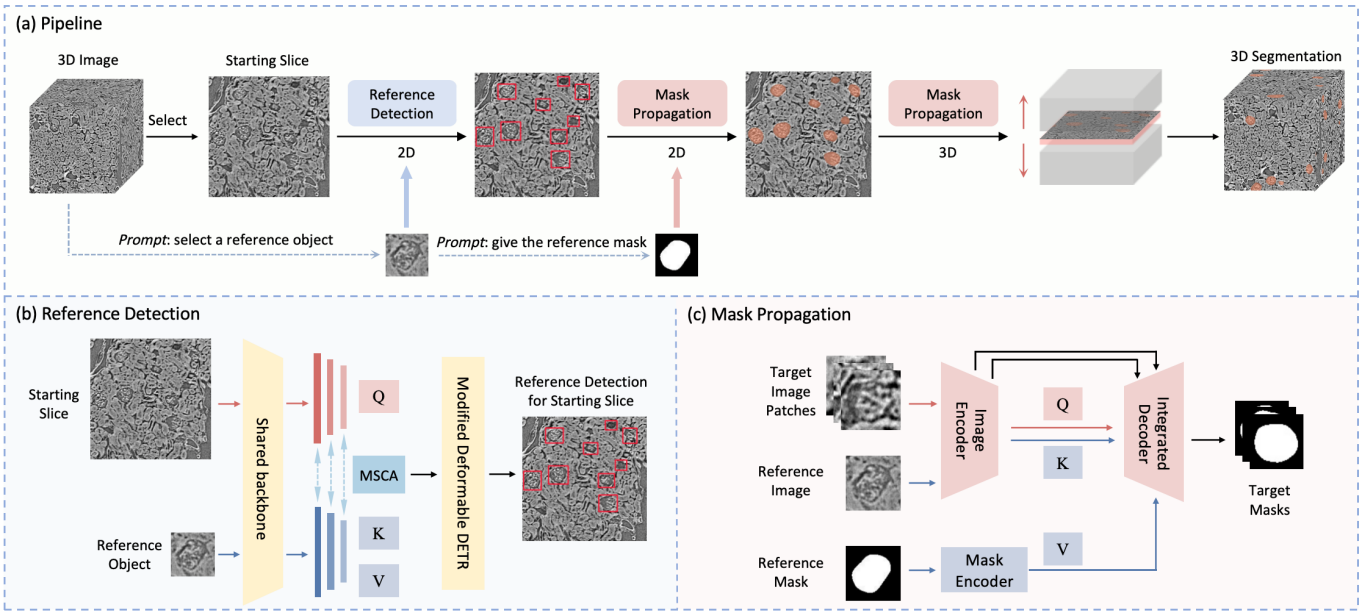

Fig. 1. An overview of the proposed RDMP. (a) The RDMP pipeline. (b) The architecture of reference detection model. (c) The architecture of mask propagation model. 2D mask propagation uses the reference image and its mask to segment detected patches within the starting slice, while 3D mask propagation extends starting slice's mask to neighboring slices, producing 3D segmentation.

detection can be corrected by 2D mask propagation, while false negatives are mitigated through 3D mask propagation.

The key contributions of our work are as follows:

- Development of a novel Reference Detection and Mask Propagation (RDMP) pipeline for segmenting large repetitive structures, ensuring consistent robustness and accuracy.
- Achievement of high accuracy in large-scale Synchrotron Radiation X-ray (SRX) volumes while requiring minimal 2D bounding box annotations for fine-tuning the model trained on microCT data, demonstrating strong generalization capability.
- Provision of a foundational algorithmic framework for segmenting massive microscopic structures in high resolution 3D imaging modalities, particularly demonstrated in microCT and SRX volumes.

2. METHODOLOGY

The RDMP methodology, illustrated in Fig. 1, consists of three sequential stages: 2D reference detection, 2D mask propagation, and 3D mask propagation.

2.1. 2D Reference Detection

Given a starting slice x_{s} and a reference object x_{r} cropped from a slice of the 3D image, the 2D reference detection aims to locate all glomeruli within x_{s}. As depicted in Fig. 1(b), a visual-prompt branch and a multi-scale cross-attention (MSCA) module are integrated into the network architecture to provide reference information and improve small-object detection. Specifically, x_{s} and x_{r} are processed through a shared backbone to generate multi-scale feature maps $\{x_{\mathrm{s}}^l\}_{l=1}^L$ and $\{x_{\mathrm{r}}^l\}_{l=1}^L$. For each feature level l, identical position encodings pos_l are added to preserve spatial context:

$$\tilde{x}_{\mathrm{s}}^l = x_{\mathrm{s}}^l + \mathrm{pos}_l,$$
$$\tilde{x}_{\mathrm{r}}^l = x_{\mathrm{r}}^l + \mathrm{pos}_l.$$

The starting slice features act as the query Q, while the reference object features serve as the key K and the value V in the cross-attention operation:

$$x_{\mathrm{f}}^l = \mathrm{CrossAttn}(\tilde{x}_{\mathrm{s}}^l, \tilde{x}_{\mathrm{r}}^l) = \mathrm{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V,$$

where $Q = W_Q \tilde{x}_{\mathrm{s}}^l$, $K = W_K \tilde{x}_{\mathrm{r}}^l$, and $V = W_V \tilde{x}_{\mathrm{r}}^l$, with W_Q, W_K, and W_V as learnable projection matrices, d scaling the dot product for computational stability. This cross-attention mechanism, applied at each feature level, effectively captures multi-scale fusion information, providing a comprehensive representation suited for detecting small structures like glomeruli. The fused features $\{x_{\mathrm{f}}^l\}_{l=1}^L$ are subsequently processed by the encoder-decoder component, based on a modified deformable DETR [11], to generate the final detection box coordinates $\{b_{\mathrm{s},k}\}_{k=1}^K$ and confidence scores $\{c_{\mathrm{s},k}\}_{k=1}^K$.

2.2. 2D Mask Propagation

After obtaining the detection boxes $\{b_{\mathrm{s},k}\}_{k=1}^{K}$ and confidence scores $\{c_{\mathrm{s},k}\}_{k=1}^{K}$ for the starting slice x_{s}, 2D mask propagation generates segmentation for boxes with confidence scores higher than a detection threshold t. Using the reference object image x_{r} and its mask m_{r}, a fine-tuned PAM model (illustrated in Fig. 1(c)) propagates mask information to each detected box $\{b_{\mathrm{s},k}\}_{k=1}^{K}$. The output individual glomerulus masks $\{m_{\mathrm{s},k}\}_{k=1}^{K}$ are then consolidated into their respective positions to create the full mask m_{s} for x_{s}.

To mitigate erroneous mask propagation in false positive detections, we incorporate both positive samples of glomeruli-to-glomeruli (both connected and non-connected) and negative samples of glomeruli-to-background in the PAM fine-tuning process. This approach ensures mask generation for true glomeruli regions while producing zero-element masks for non-glomeruli regions. By selectively propagating the reference mask only to bounding boxes containing actual glomeruli, this decoupled detection-segmentation approach enhances pipeline robustness. It ensures that even with low detection thresholds generating numerous false positives, the 2D mask propagation maintains accuracy by preventing mask propagation to irrelevant areas.

2.3. 3D Mask Propagation

After obtaining the starting slice mask m_{s} through 2D mask propagation, we perform 3D mask propagation to achieve full segmentation across the entire volume. In this stage, the PAM model fine-tuned on 3D data extends the mask information from the starting slice to its $2Z$ neighboring slices along the z-axis. Using the starting slice x_{s} and its mask m_{s} as propagation prompts, the model generates segmentation results $m_{\mathrm{s}\pm Z}$ for adjacent slices $x_{\mathrm{s}\pm Z}$ through a bidirectional propagation approach. This up-down propagation strategy effectively compensates for false negatives from the 2D detection stage, particularly for initially undetected glomeruli instances. The approach ensures that even if a glomerulus is missed in one slice, it can be recovered through propagation from neighboring slices.

3. EXPERIMENTS

3.1. Datasets and Pre-processing

Our evaluation utilizes two datasets: a private microCT dataset (26 volumes of size $512 \times 512 \times 500$ and 1 volume of size $1024 \times 1024 \times 1000$) and a public SRX dataset (1 large-scale volume of size $3840 \times 3072 \times 256$) [12]. Both datasets are isotropic with an approximate resolution of $1\mu m$ and annotated by experts. Data augmentation includes normalization, random flipping, resizing, cropping, and contrast adjustment based on the dynamic range of the foreground.

For the microCT dataset, 19 samples of $512 \times 512 \times 500$ are randomly chosen for training, and the remaining samples are used for testing. For the training of 2D reference detection and 2D mask propagation, slices are extracted from the 3D volume along z-axis at intervals of 5. For each 2D slice, one glomerulus is randomly selected as a reference object. In mask propagation, glomeruli larger than 50 pixels are selected; each glomerulus g_i generates positive pairs $\{g_i, m_i, g_j, m_j\}$ by applying its mask m_i to other glomeruli g_j within the same slice. Negative samples are formed by applying m_i to random 64×64 background regions, with a 3:1 ratio of positive to negative samples. The 3D mask propagation model follows PAM's original configuration.

For the SRX dataset, we crop 30 2D images of size 512×512 from the full-kidney volume, and use these slices with bounding box annotations to fine-tune the detection model pre-trained on the microCT dataset.

3.2. Implementation Details

Our 2D reference detection model employs ResNet-50 [13] as the shared backbone, initialized with ImageNet [14] pre-trained weights and trained for 200 epochs on the microCT dataset. Other settings align with those in Deformable DETR. The 2D and 3D mask propagation models, initialized with PAM parameters, are fine-tuned for 30 and 3 epochs respectively, using the AdamW optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with batch sizes of 512 and 128, and the soft dice loss function. Given the localized nature of glomerular structures, we systematically select starting slices at fixed intervals ($z_{\text{interval}} = 30$) along the z-axis, corresponding to the approximate average glomerulus radius. All starting slices can use a shared reference object cropped from the 3D image, and the 3D mask propagation parameter Z is set to 20. To improve efficiency, volumes are divided into overlapping $256 \times 256 \times 250$ sub-volumes. Models are trained on four NVIDIA A800-SXM4-80GB GPUs using PyTorch 1.13.1 [15], with inference on one such GPU.

For detection model evaluation, only boxes with Intersection over Union (IoU) ≥ 0.5 are counted as true positives. Precision, recall, and their harmonic mean (F1 score) are calculated at different detection thresholds. The mask propagation models are evaluated using the Dice Similarity Coefficient (DSC). 2D DSC is the average score across all reference slices, while 3D DSC measures alignment between predictions and ground truth for the complete 3D volume.

3.3. Results on microCT Dataset

Tab. 1 presents quantitative evaluation results for each RDMP component, illustrating performance variations across different detection thresholds t. As t increases, precision shows steady improvement, reaching 0.9817 at $t = 0.9$, while recall exhibits a corresponding decline. This pattern reflects the in-

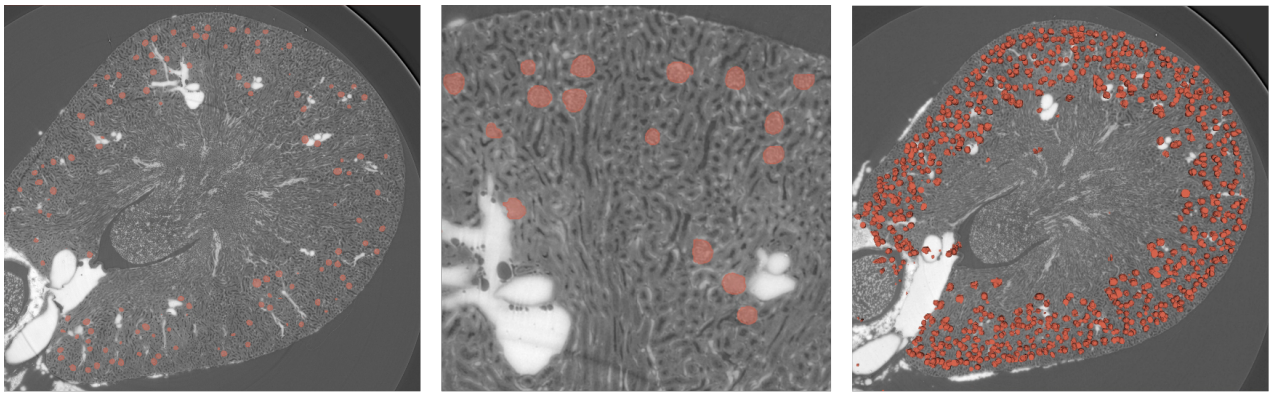

Fig. 2. Visualization of inference result on SRX of size $3840 \times 3072 \times 256$: 2D whole slide image (left), locally enlarged of 2D slice (middle), 3D visualization (right).

t	Precision	Recall	F1 Score	2D DSC	3D DSC
0.1	0.7427	**0.8475**	0.7845	**0.7051**	0.9051
0.2	0.8663	<u>0.8247</u>	0.8402	<u>0.6999</u>	0.9077
0.3	0.9051	0.8154	0.8525	0.6990	0.9088
0.4	0.9284	0.8033	**0.8560**	0.6954	**0.9096**
0.5	0.9425	0.7900	<u>0.8535</u>	0.6902	<u>0.9094</u>
0.6	0.9507	0.7779	0.8497	0.6862	0.9089
0.7	0.9615	0.7615	0.8432	0.6811	0.9086
0.8	<u>0.9714</u>	0.7417	0.8339	0.6733	0.9093
0.9	**0.9817**	0.7078	0.8141	0.6556	0.9062

Table 1. Model performance under different detection thresholds t. **Bold** indicates the best performance. <u>Underline</u> indicates the second-best performance.

herent trade-off between maximizing true positives and minimizing false positives. The F1 score, balancing precision and recall, peaks at 0.8560 with $t = 0.4$, indicating optimal detection model performance. At this threshold, the 3D DSC score also reaches its maximum of 0.9096, indicating how precise detection enhances overall pipeline performance.

Our results demonstrate effective mitigation of both false positives and false negatives introduced by the detection model. The 2D DSC score peaks at $t = 0.1$ and gradually declines as t increases, suggesting successful false positive suppression through 2D mask propagation. Even with numerous detected boxes, the method avoids erroneous propagation in non-glomeruli regions. Meanwhile, the 3D DSC score remains relatively stable across different values of t, reflecting that 3D mask propagation compensates for false negatives in prior stages. Missed glomeruli in individual slices can be recovered through vertical propagation from adjacent slices, ensuring volumetric segmentation continuity. This stability highlights the pipeline's robustness to detection results.

3.4. Generalize to Large-volume SRX dataset

For inference on the SRX volume with full horizontal coverage, 30 512×512 2D images with bounding box annotations, as described in 3.1, are used to fine-tune the detection model pre-trained on the microCT dataset. We directly apply the microCT-trained 2D and 3D mask propagation models without additional fine-tuning. This configuration achieves a 3D DSC score of 0.8431 on the large-volume SRX dataset, with results visualized in Fig. 2.

The results demonstrate successful detection and segmentation of most glomeruli in the large, cross-modal volume, despite limited fine-tuning data. This performance validates the adaptability and efficiency of our pipeline in cross-modal applications, establishing it as a resource-efficient solution for comprehensive kidney analysis.

4. CONCLUSION

In this work, we propose a Reference Detection and Mask Propagation (RDMP) pipeline for efficient 3D glomeruli segmentation, including 2D reference detection, 2D mask propagation, and 3D mask propagation. Experiments on microCT and SRX datasets demonstrate that our pipeline is robust to detection errors and achieves high segmentation accuracy while requiring minimal bounding box annotations for fine-tuning. The success of RDMP in processing high-resolution volumetric data, particularly its strong performance on cross-modal datasets, establishes it as a promising framework for analyzing large-scale repetitive structures in biomedical imaging. This capability is especially valuable for advancing quantitative analysis in kidney research and broader biomedical applications where efficient processing of massive microscopic structures is essential.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Peking University First Hospital (approval number: J2022138).

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (12090022 to B.D., 11831002 to B.D., 81801778 to L.Z.), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041 to L.Z.).

7. REFERENCES

[1] Anusha Aswath, Ahmad Alsahaf, Ben NG Giepmans, and George Azzopardi, "Segmentation in large-scale cellular electron microscopy with deep learning: A literature survey," *Medical image analysis*, p. 102920, 2023.

[2] Rui Santos, Max Bürgi, José María Mateos, Alessandro Luciani, and Johannes Loffing, "Too bright for 2 dimensions: recent progress in advanced 3-dimensional microscopy of the kidney," *Kidney International*, vol. 102, no. 6, pp. 1238–1246, 2022.

[3] Luke Xie, Georgios Koukos, Kai Barck, Oded Foreman, Wyne P Lee, Robert Brendza, Jeff Eastham-Anderson, Brent S McKenzie, Andrew Peterson, and Richard AD Carano, "Micro-ct imaging and structural analysis of glomeruli in a model of adriamycin-induced nephropathy," *American Journal of Physiology-Renal Physiology*, 2019.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[5] Tausif Diwan, G Anirudh, and Jitendra V Tembhurne, "Object detection using yolo: Challenges, architectural successors, datasets and applications," *multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, 2023.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[7] Qing Jiang, Feng Li, Tianhe Ren, Shilong Liu, Zhaoyang Zeng, Kent Yu, and Lei Zhang, "T-rex: Counting by visual prompting," *arXiv preprint arXiv:2311.13596*, 2023.

[8] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang, "T-rex2: Towards generic object detection via text-visual prompt synergy," *arXiv preprint arXiv:2403.14610*, 2024.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[10] Zifan Chen, Xinyu Nan, Jiazheng Li, Jie Zhao, Haifeng Li, Ziling Lin, Haoshen Li, Heyun Chen, Yiting Liu, Lei Tang, Li Zhang, and Bin Dong, "Pam: A propagation-based model for segmenting any 3d objects across multi-modal medical images," *arXiv preprint arXiv:2408.13836*, 2024.

[11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[12] Willy Kuo, Diego Rossinelli, Georg Schulz, Roland H Wenger, Simone Hieber, Bert Müller, and Vartan Kurtcuoglu, "Terabyte-scale supervised 3d training and benchmarking dataset of the mouse kidney," *Scientific data*, vol. 10, no. 1, pp. 510, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.