Advances in kidney biopsy lesion assessment through dense instance segmentation

Xiong, Z.; He, J.L.; Valkema, P.; Nguyen, T.Q.; Naesens, M.; Kers, J.; Verbeek, F.J.

Artificial Intelligence In Medicine 164 (2025) 103111

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Advances in kidney biopsy lesion assessment through dense instance segmentation

Zhan Xiong [a], Junling He [b], Pieter Valkema [c], Tri Q. Nguyen [d], Maarten Naesens [e,f], Jesper Kers [b,c,g,1,2,3], Fons J. Verbeek [a,*,3]

[a] LIACS, Leiden University, Snellius Gebouw, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands
[b] Department of Pathology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands
[c] Department of Pathology, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands
[d] Department of Pathology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands
[e] Department of Nephrology and Renal Transplantation, University Hospitals Leuven, Herestraat 49, 3000, Leuven, Belgium
[f] Department of Microbiology, Immunology, and Transplantation, KU Leuven, Oude Markt 13, 3000, Leuven, Belgium
[g] Van't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands

ARTICLE INFO

MSC:
68T07
68T45
65D15
65D18

Keywords:
Renal pathology
Dense instance segmentation
Diffusion model
Regional transformers
Multi-label lesion classification

ABSTRACT

Renal biopsies are the gold standard for the diagnosis of kidney diseases. Lesion scores made by renal pathologists are semi-quantitative and exhibit high inter-observer variability. Automating lesion classification within segmented anatomical structures can provide decision support in quantification analysis, thereby reducing inter-observer variability. Nevertheless, classifying lesions in regions-of-interest (ROIs) is clinically challenging due to (a) a large amount of densely packed anatomical objects, (b) class imbalance across different compartments (at least 3), (c) significant variation in size and shape of anatomical objects and (d) the presence of multi-label lesions per anatomical structure. Existing models cannot address these complexities in an efficient and generic manner. This paper presents an analysis for a **generalized solution** to datasets from various sources (pathology departments) with different types of lesions. Our approach utilizes two sub-networks: dense instance segmentation and lesion classification. We introduce **DiffRegFormer**, an end-to-end dense instance segmentation sub-network designed for multi-class, multi-scale objects within ROIs. Combining diffusion models, transformers, and RCNNs, DiffRegFormer is a computational-friendly framework that can efficiently recognize over 500 objects across three anatomical classes, i.e., glomeruli, tubuli, and arteries, within ROIs. In a dataset of 303 ROIs from 148 Jones' silver-stained renal Whole Slide Images (WSIs), our approach outperforms previous methods, achieving an Average Precision of 52.1% (detection) and 46.8% (segmentation). Moreover, our lesion classification sub-network achieves 89.2% precision and 64.6% recall on 21889 object patches out of the 303 ROIs. Lastly, our model demonstrates direct domain transfer to PAS-stained renal WSIs without fine-tuning.

1. Introduction

The evaluation of expert pathologists of kidney biopsies remains the gold standard for diagnosing and staging renal diseases [1]. Although biopsies digitalized into Whole Slide Images (WSIs)[4] have facilitated obtaining a visual morphological assessment of different anatomical structures for disease categorization, high-quality diagnostic assessments heavily depend on the correct lesion quantification manually annotated by pathologists across structures within a biopsy. Fig. 1 shows an example of annotated region-of-interest (ROI)[5] within a biopsy that contains hundreds of densely packed tissue objects. The

* Corresponding author.
E-mail address: f.j.verbeek@liacs.leidenuniv.nl (F.J. Verbeek).

[1] Data and annotations can be made available upon reasonable request due to the cooperation with multiple hospitals. Please email to J. Kers, MD,Ph.D.: j.kers@amsterdamumc.nl

[2] Source code is available on request from LIACS GitLab under the Apache 2.0 license. Please email to Prof. Fons Verbeek: f.j.verbeek@liacs.leidenuniv.nl

[3] Shared senior authorship.

[4] https://www.mbfbioscience.com/whole-slide-imaging-analysis/.

[5] To prevent any confusion with 'RoI' in RCNNs, we will hereafter use the term 'ROI' solely in the context of renal biopsies. To maintain clear distinctions in RCNN discussions, we will adopt the terms 'candidate regions' or 'regional proposals' s.

https://doi.org/10.1016/j.artmed.2025.103111
Received 21 May 2024; Received in revised form 6 February 2025; Accepted 14 March 2025
Available online 23 March 2025

annotation would cost a skilled expert around 2–4 h for a complete biopsy. Due to the complexity and time-consuming nature of this task, there is a strong need for automated structure annotation and tools for lesion classification to facilitate further quantification, offload annotation time, and reduce intra- / inter-observer variability [2,3].

Deep learning based instance segmentation algorithms have demonstrated significant capabilities on biomedical datasets [4–12]. However, developing a generic framework in renal pathology is still a challenge. This challenge can be elaborated in four technical gaps and two practical issues. The four gaps (see Fig. 1) are: (1) densely packed structures, up to 1000, per ROI; (2) considerable variation in size and shape of objects (e.g., arteries can be up to 100 times larger than that of tubuli); (3) class imbalance (e.g., the tubulointerstitial area occupies more than 70% on average in healthy and diseased renal parenchyma [5]); (4) each anatomical structure may present multiple lesions. The two practical issues are: (a) how to fuse multiple datasets with variation in staining to fully exploit scarce annotations; (b) readiness for extensibility, i.e., cost-effective adaptation to new lesion types from expanding datasets in clinical scenarios. Simultaneously addressing these difficulties requires a universal framework that includes efficiency, staining style (domain) transfer [13–15], and flexibility for continuous learning [16,17].

Prior studies have shown significant capabilities in lesion classification through dense instance segmentation, but their paradigms lack scalability and adaptability to datasets with potential changes in lesion compositions. Some approaches [5,7,8,18,19] adopted a two-step process with semantic segmentation followed by dataset-specific post-processing to achieve final instance masks, which cannot be scaled to large-scale datasets with dense objects in various shapes. Besides, each lesion is defined as one semantic class, leading to difficulty for multi-label lesion classification and potential changes in lesion combinations. Specifically, adding or removing lesions requires complete redesign and retraining of the model. Lastly, increasing lesion types requires adding more segmentation maps for prediction, which is significantly resource-intensive.

Detection-based models with regional convolution neural networks (RCNNs) [6,20] can efficiently detect dense anatomical structures and separately design lesion classification heads for each class. They can, therefore, adopt a plug-and-play mechanism and adapt to lesion changes by replacing the corresponding sub-modules and maximally reusing the others. However, these variants are unscalable to multi-class objects with various scales and shapes due to the reliance on pre-defined bounding box anchors, limiting their application to process a single class with lesser variation in scales and shapes, e.g., glomeruli.

Recently, transformers have been the prevalent anchor-free approach to address multi-class objects at various scales and shapes. Transformer-based instance segmentation utilizes attention mechanisms and learns latent representative embeddings (queries) from global contextual features to process vastly varying objects. However, existing models [21–23] are resource-intensive for dense structures in large-scale datasets. That is due to two shortages: (1) a large number of static queries from one embedding per object; (2) low instance map occupancy from one object per instance map (depicted in Fig. 2c). Hence, they are limited to processing classes with sparse objects, like glomeruli.

In response to these challenges, we propose a generic and extensible system for dense instance segmentation and lesion classification on large-scale datasets with potential changes in lesion combinations. Our design has two key components: (a) A novel dense instance segmentation sub-network that recognizes basic anatomical structures, i.e., glomeruli, tubuli, and arteries; (b) a lesion classification sub-network with a set of independent heads that predicts lesions for each class. It is crucial to separately modularize lesion classification and dense instance segmentation for the big picture. The segmentation of basic anatomical structures is a universal foundation for all visual assessment systems. However, lesion classification is a task-driven

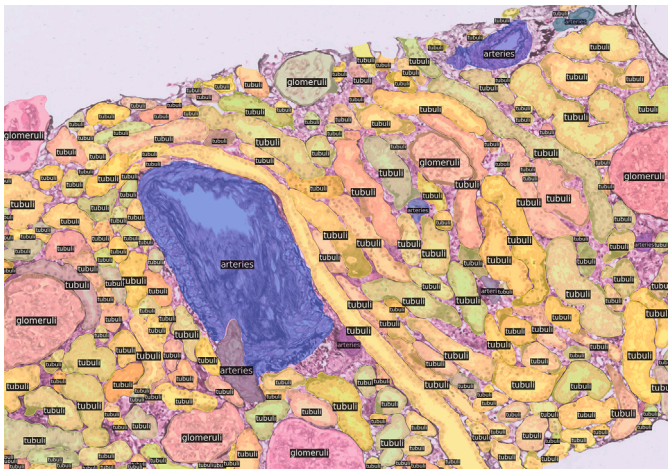


Fig. 1. An illustration depicts a manually annotated ROI of kidney biopsy, focusing on glomeruli, tubuli, and arteries. There are three primary challenges encountered in clinical renal biopsies: (1) a large number of objects closely touching each other; (2) the significant variation in size and shape among different instances; (3) the distribution of classes is heavily biased.

downstream application, which might vary depending on the interest of users. Therefore, the robustness of dense instance segmentation from local changes in heads of lesion classification is beneficial for the extensibility of our system and for expanding datasets in clinical scenarios.

More specifically, for dense instance segmentation, we propose a novel end-to-end approach, named **DiffRegFormer**, which effectively combines the advantages of diffusion models, RCNNs, and transformers to tackle dense objects with multi-class and multi-scale. Diffusion models [24] generate bounding box proposals from Gaussian noise, and therefore eliminating the need for pre-defined anchors; RCNNs efficiently crop dense instance maps into dense regions which have very high occupancy rate w.r.t. the bounding boxes (see Fig. 2. (a)); Cross-attention mechanisms with dynamic queries [25–29] extract long-range contextual features and enables robust representation of objects across varying scales. DiffRegFormer is not just a simple assembly of the aforementioned techniques. Such assemblies may suffice for sparse instance segmentation on datasets like MSCOCO [30]. They, however, fail in the context of dense instance segmentation of kidney biopsy ROIs. The crux roots in the proposals that are essentially Gaussian noise at the early training stage of diffusion models. Those noisy candidates, consequently, hinder training due to accumulating errors. Our ablation study analyzes these effects in detail (see Section 4.4.1) and highlights the importance of our specific designs. Specifically, we introduce the following **key innovations** to address the challenges for dense instance segmentation:

- **Regional features**: Similar to RCNNs, feature maps are cropped using generated proposals and converted to dynamic queries for efficient long-range dependency modeling; this is robust to large-scale variations.
- **Feature disentanglement**: Instead of using shared feature maps, we redesign the model structure to generate separate feature maps for the bounding box decoder and mask decoder; this is crucial to stabilize the training of the mask decoder.
- **Class-wise balanced sampling**: Unlike conventional sampling methods, we propose a novel sampling approach. The key difference is to select **class-wise balanced** positive samples among the **ground-truth** boxes instead of the proposal ones.

In conclusion, based on those key designs, the proposed model is a generic and extensible approach for large-scale datasets with little

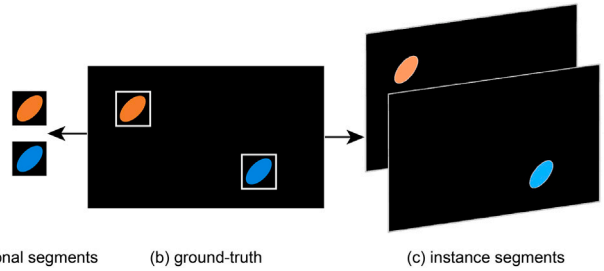


Fig. 2. Comparison between regional segments and instance segments. (a) cropped instance maps from bounding boxes tightly surrounding each object; (b) ground-truth bounding boxes and instance segments within one image; (c) separated entire instance map per object.

overhead. To the best of our knowledge, DiffRegFormer is the first end-to-end framework that combines diffusion methodology with a transformer in RCNN-style to process dense objects within ROIs efficiently for multi-scale and multi-class. Fig. 3 depicts a flow chart of the dense instance segmentation sub-network.

Each anatomical structure has a dedicated classification head for the lesion classifier since class-wise lesions can differ depending on the dataset's lesion combinations. In Fig. 6, we input the croppings of anatomical structures and resized them to the same size. Each head is trained only with patches of the specific structures and learns to predict multi-label lesions over each class. Our dataset currently only contains sclerotic glomeruli and atrophic tubuli. However, the lesion classifier is extensible to the expanded large-scale dataset at minimal cost because adding, removing, or replacing heads cannot disrupt the other modules.

Our work aims to propose a **generic technical solution** that has flexibility and scalability for automating lesion classification on dense anatomical structures in clinical scenarios. It works as the foundation for further quantification analyses. The main contributions of this work are the following:

- We propose the first end-to-end dense instance segmentation that effectively combines diffusion models, RCNNs, and transformers for multi-class, multi-scale objects. More importantly, our model can directly process ROIs of kidney biopsies, avoiding patch splitting.
- We propose novel designs to address the accumulative errors caused by diffusion models at an early stage and stabilize the training process.
- We propose a novel class-wise balance sampling method to improve detection and segmentation performance.
- Instead of the static queries used in transformers, we convert regional features into dynamic queries and model the long-range dependencies between regional features while avoiding performance deterioration on dense objects.
- We compare DiffRegFormer with previous models that can process multi-class and multi-scale objects within ROIs in an end-to-end manner. Our model outperforms the previously published models in evaluating Jones' silver-stained images.
- We show that DiffRegFormer has the potential of stain (domain)-agnostic detection for PAS-stained images without stain-specific fine-tuning.
- Our lesion classifier can achieve multi-lesion classification. In addition, our plug-and-play strategy can flexibly adapt to large-scale datasets with potential changes in lesion combinations at minimal overhead.

The remainder of the paper is organized as follows. In Section 2, all relevant work is introduced, and we elaborate on all improvements w.r.t. previous research. In Section 3, we describe each component of our model in detail. In Section 4, we demonstrate the extensive evaluation of our model, including comparison experiments and ablation studies. In Section 5, we describe the advantages and limitations of our pipeline and show possible future follow-up research.

2. Related work

Semantic Segmentation with Post-Processing: In the context of lesion classification of dense structures for renal biopsies, semantic segmentation with post-processing has drawn substantial attention. Previous studies [5,7,8,18] treat anatomy-wise lesions (e.g., atrophic tubuli) as semantic classes alongside basic anatomical structures (e.g., arteries, glomeruli). An auxiliary **border** class is often integrated during the training phase to facilitate splitting semantic masks into distinct instances. This approach has demonstrated notable performance. However, challenges are faced in scalability and adaptability to large-scale datasets with potential changes in lesion composition. These issues are at hand as (1) Reliance on ad-hoc procedures (e.g., thresholding and morphological operations [31]) for border generation reduces generalizability. (2) Due to the inflexibility of dataset change, any modification to lesion types necessitates model redesign and retraining. (3) Computational cost scales with lesion classes in the dataset. (4) The difficulty in modeling long-range correlations between objects of varying size due to splitting ROIs into tiles. (5) It cannot tackle multi-label lesion classification. In contrast to these issues, our end-to-end framework overcomes these limitations by directly modeling dense objects within ROIs through attention mechanisms.

Diffusion Model: Diffusion models, a class of deep generative models [24,32], learn to approximate complex distributions through iterative denoising. Despite success in data generation [33–35], natural language processing [36], audio processing [37], and self/weakly-supervised learning [38], their application to dense instance segmentation remains limited. Initial attempts have been to introduce diffusion models to segmentation tasks [39,40]. The adaptation to dense instance segmentation, however, remains challenging. From our ablation studies, we formulate two primary challenges: (1) The potential dominance of the diffusion modules over feature learning in shared feature maps destabilizes the training of instance prediction modules. (2) Accumulative errors from the diffusion module's early training stages impair the accuracy of dense object predictions. Our approach introduces innovative strategies to overcome these challenges, enabling the first successful implementation of a diffusion model for dense instance segmentation.

Transformer: Transformer-based approaches introduce an anchor-free paradigm in an end-to-end manner for instance segmentation, leveraging their capability to capture long-range dependencies and aggregate contextual information. In general, the dependency extraction process is called the attention mechanism, while each context feature is denoted as a query for object representation. Early transformer-based methods [21–23] rely on static queries to model all objects within a dataset. These approaches are incompatible with dense object scenarios due to the inevitable expansion of the static query set with dataset size. Recent works show [26–29] advancement towards dynamic queries, allowing for temporary modeling of objects within a mini-batch. In this manner, the complexity of processing dense objects is reduced. However, their instance segment representation remains inefficient for dense instance segmentation[6] due to the expensive representation of instance segments (Fig. 2.(c)) with low occupation rate. Our DiffRegFormer addresses

[6] We conducted experiments on our dataset from the official code of both. Compared to our approach, they required more than twice the GPU memory with a batch of size 2 with 500 queries. That is not infeasible for applications with dense objects.

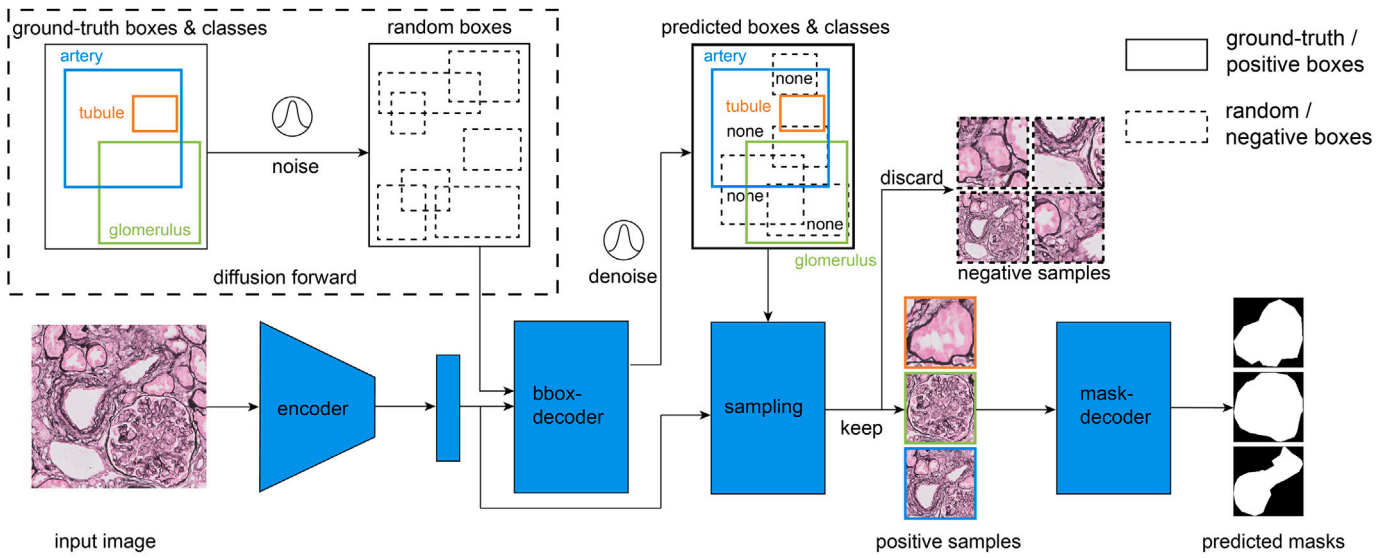


Fig. 3. Our DiffRegFormer is a one-stage anchor-free method. Instead of pre-defined anchors, we impose Gaussian noise on ground-truth boxes and generate a fixed-sized set of random bounding boxes (bbox). With feature maps extracted from the encoder, the bbox-decoder iteratively learns to denoise and predicts class-wise candidate boxes. Due to many unevenly distributed candidates, we propose a sampling module that effectively discards negative samples while maintaining a proportion of balanced positive samples for fast convergence. Finally, we make final instance masks according to the selected positive samples in the mask-decoder. For simplicity in illustration, we only choose one anatomical object per class (artery, tubule, glomerulus).

this by integrating region proposals with transformers for improved representation efficiency.

RCNN: The Regional Convolutional Neural Network (RCNN) framework [41,42] effectively generates candidate regions (i.e. proposals) for dense objects. For instance segmentation, RCNN-based methods used mask modules to predict binary instance masks within proposals. Early variants [6,20,43,44] relied on class-agnostic proposals followed by class-specific mask prediction, constrained by the limitations of pre-defined anchors. More recent approaches [45] have explored the mapping from proposals to static queries, employing cross-attention mechanisms for instance mask prediction. However, this paradigm faces scalability issues in dense object segmentation due to the computational overhead associated with static queries (see QueryInst in Table 1). Our method adopts dynamic queries, balancing computational efficiency and scalability for dense instance segmentation.

In summary, our work combines diffusion modeling and attention mechanisms into an RCNN-style framework, which addresses challenges of scalability, adaptability, and efficiency inherent in existing approaches for dense instance segmentation in renal biopsies.

3. Methods

3.1. Preliminaries

Diffusion model. Diffusion approaches [46,47] is a family of deep generative models inspired by the principles of non-equilibrium thermodynamics [33,48]. Their operation is conceptualized as T-step sequential process, iteratively transiting from an initial state Z_0 to a final state of pure noise Z_T. Diffusion forward involves the progressive addition of Gaussian noise at each transition. Due to the special properties of the Gaussian distribution [24], it is possible to sample any intermediate noisy at state Z_t directly:

$$q(Z_t|Z_0) = \mathcal{N}(Z_t|\sqrt{\bar{\alpha}_t}Z_0, (1-\bar{\alpha}_t\mathbf{I})) \tag{1}$$

where Z_0 is the original data, Z_t ($t \le T$) is the latent state at step t, and $\bar{\alpha}_t$ is the cumulative noise variance schedule [39]. Notably, the diffusion forward process does not introduce trainable parameters. The objective of a diffusion model is to learn the reverse process. A neural network, $f_\theta(Z_t, t)$, approximates the recovery of the original distribution Z_0 from a noisy state Z_t. Training optimizes this network by minimizing a loss function to denoise at arbitrary intermediate state t:

$$\mathcal{L} = \frac{1}{2}\|f_\theta(Z_t, t) - Z_0\|^2 \tag{2}$$

where Z_0 represents the ground-truth and Z_t is the noisy input. During inference, the model f_θ reconstructs from pure noise Z_T to the predicted data Z_0 through a series of reverse updates [46] with a step-size s: $Z_T \to Z_{T-s} \to \cdots \to Z_0$. See [46] for a detailed derivation.

3.2. Dense instance segmentation model

The iterative proposal generation in DiffRegFormer necessitates multiple model executions at inference. That poses a computational burden if the entire model runs on the raw image at each step during training. To address this issue, we decouple the image encoder from the remaining modules. This strategy enables the encoder to extract multi-scale feature maps from the raw input image only once. Then, subsequent modules operate for iterative refinement of box and mask predictions from an initial state of Gaussian noise conditioned on these deep features. Moreover, to facilitate the mask-decoder training, we introduce a novel sampling module that randomly selects positive samples from each class, ensuring a balanced distribution of class-wise regional features. That facilitates learning object representations with lower frequencies in the dataset.

In summary, the DiffRegFormer comprises four components:

- **Encoder**: Processes the raw input image once to extract multi-scale feature maps for bbox-decoder and mask-decoder separately.
- **Bbox-decoder**: Utilizes the extracted feature maps to iteratively refine bounding box predictions.
- **Sampling Module**: Stabilizes the training by randomly selecting class-wise balanced positive samples for the mask-decoder.
- **Mask-decoder**: Iteratively refines mask predictions based on the sampled regional features.

3.2.1. Encoder

The image encoder extracts high-level feature maps from the raw image for subsequent decoders. To generate multi-scale feature maps,

it utilizes a ResNet [49] pre-trained on ImageNet [50] as its backbone, followed by a Feature Pyramid Network (FPN) [51]. After FPN, a stack of 3 × 3 convolutional operations is applied to produce two separate feature maps, thereby decoupling the bbox-decoder and mask-decoder.

3.2.2. Bbox-decoder

The bbox-decoder builds upon DiffusionDet [39]. As shown in Fig. 4.(a), a set of random boxes and multi-scaled feature maps is first taken from the encoder as input. Then, dynamic queries (Fig. 4b) are initialized from a regional pooling operator (RoIAlign [43]) and a feed-forward network (FFN) [52]. These queries undergo iterative refinement through multiple stages. Each stage takes proposal boxes and dynamic queries from the previous stage, generating refined dynamic queries, box predictions, and class predictions for the next stage. Fig. 4c illustrates one detailed refinement stage. First, new regional features (cropped by RoIAlign) interact with dynamic queries (after a self-attention module) via dynamic convolution [53], emphasizing regions likely containing objects while suppressing others. Next, the enhanced regional features are processed by FFNs to produce refined box predictions, box-wise classifications, and queries, respectively.

3.2.3. Sampling

Conventional RCNN models use intersection over union (IoU) to classify a proposal box as positive if its IoU with a ground-truth box is ≥0.5 [43]. This approach, however, is ill-suited for diffusion methods, particularly in early training stages when diffusion methods predominantly produce noise. That can complicate convergence or even cause failure due to error accumulation. Further, conventional IoU-based methods are susceptible to class imbalances, biasing the learning process towards frequently occurring objects, thereby neglecting rare instances. We, therefore, propose a novel sampling method to address these issues. We first replace proposal boxes with ground-truth boxes, ensuring a sufficient supply of positive samples. That is justified because the bbox-decoder aims to predict boxes that closely approximate the ground-truth. We then divide ground-truth bounding boxes into groups based on their classes. Within each group, we randomly select up to n boxes as positive samples without repetition. The core idea is that objects of each class having similar shapes and appearances can be mapped to a cluster. Consequently, it allows us to learn instance feature representations with a small sample number n. This especially applies to the tubular class that occupies most of our dataset. This sampling strategy offers three distinct advantages:

1. **Rare Instance Learning**: The mask-decoder can effectively learn representations of infrequently occurring instances whenever they appear in the input, mitigating the risk of being overwhelmed by more common classes.
2. **Randomize Frequent Instance**: Learning on a randomized subset of frequent instances prevents model over-fitting.
3. **Training Stability**: Employing selected ground-truth boxes as positive samples promotes stability and accelerates mask training.

3.2.4. Mask decoder

The mask decoder iteratively refines mask predictions, utilizing regional features cropped from multi-scale feature maps based on boxes selected by the sampling module. Instead of using static queries [22, 54], we compute cross-attention between dynamic queries and positive regional features. That highlights effective queries and filters others, enhancing regional features by focusing on pixels within positive regions, thereby facilitating accurate per-pixel instance score prediction (Fig. 5).

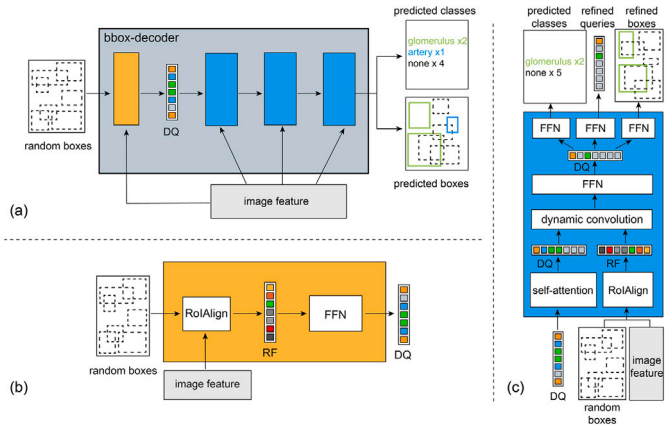


Fig. 4. The bounding box (bbox)-decoder takes multi-scale feature maps and a set of random boxes as input. Then, the prediction of classes and boxes will be outputted iteratively. (a) The module comprises an initialization head for dynamic queries (orange rectangle) and multiple box refinement heads (blue rectangles). (b) In the orange rectangle in (a), the initial dynamic queries are generated via a RoIAlign pooling operator and a feed-forward network (FFN). (c) Each box refinement module (one blue rectangle in (a)) takes the previous stage's dynamic queries and proposal boxes as input, generating predictions and refined dynamic queries for the next stage. Abbreviations: dynamic queries: DQ, feed-forward network: FFN, regional features: RF. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

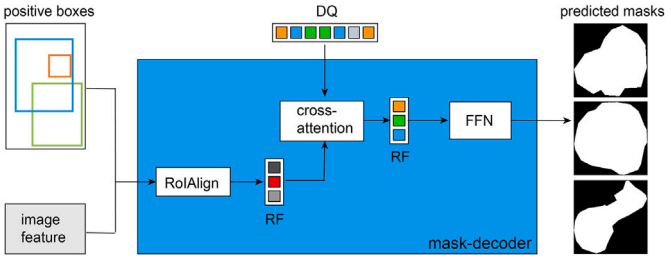


Fig. 5. The mask-decoder takes multi-scale feature maps and a set of positive boxes as input and predicts instance masks. The dynamic queries interact with regional features using cross-attention and only highlight pixels residing in proposal boxes. Final instance masks are generated from the enhanced regional feature maps. Abbreviations: dynamic queries: DQ, feed-forward network: FFN, regional features: RF.

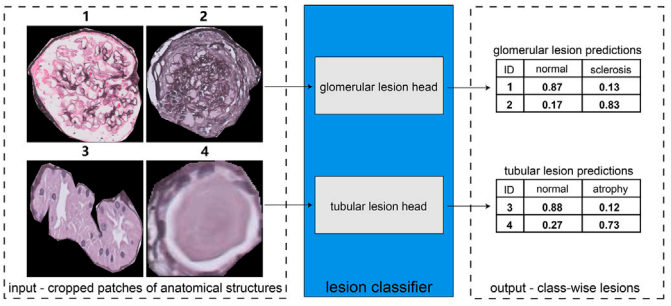


Fig. 6. The lesion classifier consists of 2 prediction heads, one dedicated to glomeruli and the other to tubuli. It takes croppings of anatomical structures and outputs the probability of class-wise lesions.

3.3. Lesion classifier

The lesion classifier comprises multiple independent lesion prediction heads. Each head specializes in a specific anatomical structure (e.g., glomerulus) and processes cropped images of the specified structure to predict probabilities for multi-label lesions. Fig. 6 illustrates this design. The architecture of a single head consists of three convolutional

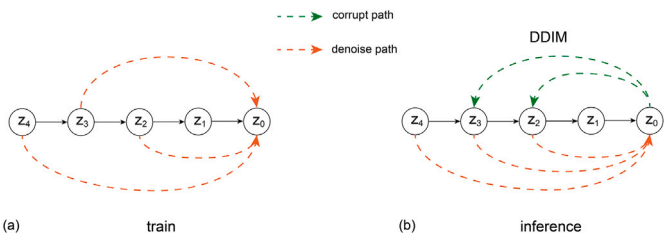


Fig. 7. An illustration for a diffusion model at different stages depicts a simplified process with 5 states (skip step = 1), where state 4 is pure Gaussian noise and state 0 is the ground-truth. (a) During training, a diffusion model randomly selects denoising paths to train its decoder; (b) In inference, the Denoising Diffusion Implicit Model (DDIM) flexibly alternates between denoising and corrupting paths, generating multiple predictions that are then ensembled for a refined final result.

blocks followed by a max-pooling operator. Each block contains a 3×3 convolutional, batch normalization, and ReLU activation. We currently implement two prediction heads for glomeruli (predicting sclerosis) and tubuli (predicting atrophy).

Notably, the modular design of the instance segmentation component allows lesion prediction heads to operate as independent plug-ins. This modularity is crucial for seamlessly adapting the model to the expanding datasets, particularly in clinical scenarios. It allows for the straightforward integration or removal of prediction heads with minimal impact on the overall model architecture. For example, if new annotations for arteries become available, a dedicated prediction head can be trained to classify arterial lesions and integrate them into the model without modifying existing components. Likewise, this design extends the model's diagnostic scope to include additional lesions within a given structure (e.g., tubuli) by simply replacing an existing prediction head with a newly trained one on expanded data. This design promotes the model's versatility and long-term adaptability within clinical scenarios where the scope of identifiable lesions may increase.

3.4. Implementation details

The entire model adopts distinct strategies for the training and inference phases. During training, the DiffRegFormer learns to reconstruct ground-truth boxes through adaptation at state 0 from an arbitrary noisy state t (Fig. 7.(a)). Additionally, the sampling module provides class-balanced ground-truth boxes to facilitate the mask decoder's training. Lastly, The lesion classifier operates on croppings generated from ground-truth instance masks. Each prediction head specializes in classifying lesions specific to the corresponding anatomical structure within these croppings. At the same time, during inference, the DiffRegFormer assembles multiple predictions to refine final results (Fig. 7b). Moreover, the sampling module is deactivated, and all proposals are directly forwarded to the mask decoder for instance segmentation prediction. Likewise, the lesion classifier utilizes instance masks predicted by DiffRegFormer to crop structures. As a last step, each prediction head predicts probabilities for class-specific lesions within these cropped regions.

3.4.1. Training

We start by constructing noisy boxes from the ground-truth. The DiffRegFormer is then trained to recover bounding boxes and iteratively generate instance masks from an arbitrary state t. The algorithm is detailed in Algorithm 1. **Diffusion forward**. Ground-truth boxes are first concatenated with additional random boxes to align to a fixed length across images, accounting for the variability in instance counts. Gaussian noise is then introduced as the noise scale governed by $\bar{\alpha}$ as per Eq. (1). Following Nichol and Dhariwal [32], the scale varies with state t using a monotonically decreasing *cosine* schedule. It should be noted that, for optimal performance, the signal-to-noise ratio requires a relatively high signal scaling value compared to image generation tasks [24,39,55–57].

Training Loss. The box decoder inputs noisy boxes and predicts category classifications and box coordinates. Given the one-to-many relationship between ground-truth and predicted boxes, first, a set-loss function is employed to match predictions to the ground-truth based on the lowest cost [21,58,59]. Furthermore, the mask decoder inputs sampled positive boxes and predicts instance masks. We use a Binary Cross Entropy loss function due to the one-to-one mapping. Finally, since multiple lesions can exist within one cropping, lesion classification adopts a Binary Cross Entropy loss function for multi-label prediction within each head.

Sampling. Our implementation uses 3 classes and a maximum of $N = 500$ objects per image. We set up to $n = \frac{500}{3} \simeq 166$ instances per class. More specifically, positive ground-truth boxes are sampled without repetition, not exceeding the class-wise maximum. Moreover, we disregard absent classes within the input. For example, if there are 2 glomeruli, 0 vessels, and 200 tubuli ground-truth boxes within an ROI, then the sampling module would select 2 glomeruli, 0 vessels, and 166 random tubuli as positive samples without duplication.

3.4.2. Inference

During inference, the DiffRegFormer starts from a standard Gaussian distribution corresponding to the final state T as defined in Eq. (1). Then, a progressive denoising operation reverses the predictions to the initial state 0. The algorithm, as depicted in Algorithm 2, outlines this process.

Inference steps. Each inference iteration involves two primary operations. First, the box decoder processes noisy boxes generated from the previous state t_{now} and predicts both categories and box coordinates. Subsequently, a Denoising Diffusion Implicit Model (DDIM) [46,60] introduces noise into the previously predicted boxes. This step generates new noisy boxes for subsequent state t_{next}, facilitating a progressive denoising process across different states. Notably, each iteration performs a single denoising operation, transitioning from state t_{now} to t_{next}. This characteristic allows for the assembly of intermediate predictions from multiple denoising steps (i.e., $t_i \rightarrow 0$ where $i \in [1, T]$) to refine the outcome. However, DiffRegFormer performs inference only once to improve efficiency, directly transitioning from pure Gaussian noise at state T to state 0 to obtain the final predictions.

Box replacement. Predicted boxes can be categorized as either ***positive*** (scores above a threshold, containing objects of interest) or ***negative*** (scores below the threshold, arbitrarily located). Directly applying negative boxes into DDIM would significantly degrade the quality of newly generated noisy boxes. This issue stems from the fact that negative boxes, originating from corrupted boxes during the training phase, significantly deviate from the Gaussian distribution. To ensure consistency between the training and inference phases, we substitute negative boxes with random boxes sampled from a Gaussian distribution.

4. Results

This section begins with an overview of the kidney biopsy dataset used in this study. We then make a fair comparison between DiffRegFormer and established end-to-end instance segmentation models specifically designed to handle dense, multi-class, and multi-scale objects at the ROI level. In addition, in order to include retrospective reflection in our analysis, we have compared our model with a semantic segmentation based method [18]. The results of this comparison are presented in Section 4.2.2 and the tables therein. All benchmark methods are reproduced utilizing the **mmdetection** package [61] to ensure consistency. Further, we evaluate the performance of our lesion classifier on cropped images of anatomical structures derived from the instance masks predicted by DiffRegFormer. Additionally, we

Algorithm 1: DiffRegFormer Training

```
def train(images, gt_boxes, gt_masks):
# images: [B, H, W, 3]
# gt_boxes: [B, *, 4]
# gt_masks: [B, H, W, G]
# B: batch size
# N: number of proposal boxes

# generate multi-scale features via encoder
feats = encoder(images)
# generate noised boxes at state t where t
# is a random integer in diffusion forward;
# Pad boxes to N with t_boxes: [B, N, 4]
t_boxes, t = diffusion(gt_boxes, mean=0, std=1)
# generate initial dynamic queries
d_query = initialize_query(t_boxes, feats)
# learn to reverse noised boxes at state t back
# to ground-truth boxes at state 0 and return
# refined dynamic queries
[pred_boxes, d_query] = box_decoder(t_boxes, feats, d_query, t)
# obtain bbox loss via the set objective function
loss_bbox = set_loss(pred_boxes, gt_boxes)
# randomly select balanced positive boxes from
# ground-truth boxes
pos_boxes = sampling(gt_boxes)
# generate predicted instance masks
pred_masks = mask_decoder(pos_boxes, feats, d_query)
# obtain mask loss via the set objective function
loss_mask = set_prediction_loss(pred_masks, gt_masks)
return loss_bbox, loss_mask
```

Algorithm 2: DiffRegFormer Inference

```
def infer(images, step, T):
# images: [B, H, W, 3]
# step: the skip length for the state transform
# T: the length of the chain
# B: batch size
# N: number of proposal boxes

# generate multi-scale features via encoder
feats = encoder(images)
# return Gaussian noise as noisy boxes at state T;
# Pad boxes to N with T_boxes: [B, N, 4]
[t_boxes, _] = diffusion(mean=0, std=1)
# generate state transform pairs skipping every step
# [(T, T-step), (T-step, T-2*step), ..., (step, 0)]
time_pairs = uniform(0, T, step)
# generate initial dynamic queries
d_query = initialize_query(t_boxes, feats)
# iterate over stages
for (t_now, t_next) iterate t_paris:
# predict boxes and dynamic query at state t_now
[pred_boxes, d_query] = box_decoder(t_boxes, feats, d_query, t_now)
# generate new noisy boxes from state t_now to t_next
t_boxes = ddim(t_boxes, pred_boxes, t_now, t_next)
# replace undesired boxes with random Gaussian noise
t_boxes = box_replace(t_boxes)
# generate predicted instance masks
pred_masks = mask_decoder(pred_boxes, feats, d_query)
return pred_boxes, pred_masks
```

have conducted a complexity analysis of DiffRegFormer to identify an optimal balance between its performance and computational burden. Finally, we present extensive ablation studies on DiffRegFormer that evaluate the impact of various training strategies.

4.1. Datasets

The biopsy specimens were prepared according to the Pathology Laboratory Protocol for kidney biopsies. The tissues were collected by

Table 1
Overall evaluation results. All models use ResNet-50 as a backbone network and have been trained 40 000 iterations on an NVIDIA GeForce RTX 3090 GPU with a batch size of 2. AP_{50} and AP_{75} denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively. AP_S, AP_M, and AP_L correspond to AP on small, medium, and large-sized objects, respectively. The best results are highlighted in bold. CMask-RCNN represents Cascade Mask-RCNN.

		Bounding boxes						Instances					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage	QueryInst_300	37.8	59.3	47.4	13.5	21.5	47.4	38.9	59.4	43.5	8.8	19.0	44.1
	QueryInst_500	30.1	31.5	22.8	0.9	8.7	24.1	32.5	31.9	22.7	0.4	7.8	27.7
Two-stage	Mask-RCNN	49.0	68.6	56.2	**18.0**	24.4	59.2	45.5	69.0	53.5	**16.4**	21.5	54.4
	CMask-RCNN	49.9	67.3	**58.4**	7.6	24.1	**61.4**	45.0	67.8	**54.5**	6.5	21.6	**54.5**
Our model	DiffRegFormer	**52.1**	**71.1**	57.7	15.4	**27.4**	**61.4**	**46.8**	**71.6**	52.8	14.1	**23.6**	54.2

Table 2
Per-class evaluation results. All models use ResNet-50 as a backbone network and have been trained 40 000 iterations on one NVIDIA GeForce RTX 3090 GPU with a batch size of 2. AP_{50} and AP_{75} denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively. AP_S, AP_M, and AP_L correspond to mAP on small, medium, and large-sized objects, respectively. The best results are highlighted in bold. CMask-RCNN represents Cascade Mask-RCNN.

		Bounding boxes - Glomeruli						Instances - Glomeruli					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage	QueryInst_300	62.6	74.5	71.6	–	–	63.7	56.1	74.5	68.1	–	–	57.1
	QueryInst_500	36.7	47.5	44.0	–	–	37.6	35.9	47.6	43.6	–	–	37.3
Two-stage	Mask-RCNN	69.4	85.8	78.6	–	–	70.4	67.0	85.8	78.8	–	–	68.1
	CMask-RCNN	70.7	84.0	81.7	–	–	72.1	65.1	84.0	79.8	–	–	66.4
Our model	DiffRegFormer	**80.9**	**94.2**	**89.8**	–	–	**80.9**	**74.3**	**94.2**	**80.9**	–	–	**74.3**
		Bounding boxes - Arteries						Instances - Arteries					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage	QueryInst_300	18.8	37.0	19.1	5.2	20.2	20.6	17.5	36.3	14.8	4.5	17.9	19.4
	QueryInst_500	4.5	10.0	2.9	0.5	5.6	4.5	4.7	10.5	3.8	0.1	5.3	5.3
Two-stage	Mask-RCNN	22.5	40.3	23.5	**20.2**	19.4	28.9	20.1	41.4	18.0	**20.2**	17.4	25.4
	CMask-RCNN	23.0	38.0	**26.9**	–	17.8	**33.3**	20.1	39.4	**21.9**	–	16.9	27.5
Our model	DiffRegFormer	**25.7**	**46.9**	25.8	13.5	**23.9**	31.6	**23.4**	**48.8**	20.2	**20.2**	**20.4**	**30.1**
		Bounding boxes - Tubuli						Instances - Tubuli					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage	QueryInst_300	45.8	66.4	51.4	**21.8**	44.3	57.9	41.2	67.3	47.6	13.1	39.0	55.8
	QueryInst_500	21.2	37.1	21.7	1.3	20.3	30.1	20.6	37.5	20.7	0.7	18.1	40.5
Two-stage	Mask-RCNN	57.5	79.6	66.4	15.8	53.8	78.2	51.4	79.9	**62.3**	12.7	47.3	**69.8**
	CMask-RCNN	58.1	79.9	66.5	15.1	**54.4**	**78.8**	51.8	80.0	61.7	13.0	**47.9**	69.6
Our model	DiffRegFormer	**58.5**	**80.2**	**67.9**	19.9	51.0	75.4	**55.8**	**80.6**	**62.3**	**16.6**	45.6	64.5

a standard procedure of operations: core needle biopsy, formalin fixed, rapidly processed by a routine tissue processor, embedded in paraffin, serially sectioned at 2–4 μm, mounted onto adhesive slides, and stained with Jones silver staining. The dataset comprises 148 patient biopsy samples (both transplant and native kidneys) collected from the multi-center archive of the Departments of Pathology at LUMC, AUMC, UMCU in the Netherlands or the archive of the Department of Nephrology at Leuven in Belgium. Before analysis, all biopsies were anonymized by a pathology staff member. Annotation was performed using the **ASAP** version 1.9 for Windows[7] and our custom software **Slidescape**.[8] The 148 Jones-stained renal WSIs were acquired from scanners at a resolution of PPM = 2 in BIG-TIFF format.[9] Openslide[10] was used to extract 303 ROIs at level 0 of the BIG-TIFF images. Table 3 outlines the dataset composition, and Fig. 8 illustrates the distribution of ROI short sizes. Furthermore, LUMC 115 PAS-stained renal WSIs were used only to test stain transfer. In particular, ROIs from the same WSIs were exclusively allocated to either training or validation. To train the lesion classifier, 21,889 patches of individual objects were extracted from 303 ROIs, including 916 glomeruli patches with binary lesion labels (382 with sclerosis and 534 normal) and 20,973 tubuli patches with binary labels (10,647 with atrophy and 10,326 normal), as detailed in Table 6.

Table 3
Origin of datasets digitized with different scanners.

Source	Scanner	WSIs	ROIs	Train	Eval
LUMC[a]	Philips UFS	65	142	130	12
AUMC	Philips UFS	42	80	74	6
UMCU	Hamamatsu XR	25	49	45	4
Leuven	Philips UFS	16	32	–	32

[a] Leiden University Medical Center.

4.2. Experiments

In our experimental analysis, we conducted a fair comparison of our proposed DiffRegFormer against well-established end-to-end instance segmentation models: Mask R-CNN [43], Cascade Mask R-CNN [62], and QueryInst [45], specifically designed to simultaneously handle dense, multi-class, multi-scale objects at the ROI level. All the models under comparison employed a ResNet-50 [49] backbone pre-trained on the ImageNet [50] and were trained on 249 ROIs over 40 000 iterations. We utilized the **mmdetection** data augmentation pipeline: each ROI was first subjected to *RandomFlip* with a 0.5 probability, followed by random selection between *RandomResize* or *RandomCrop* also with a probability 0.5. *RandomResize* adjusted the ROI to a size with the

[7] https://computationalpathologygroup.github.io/ASAP/.
[8] https://github.com/amspath/slidescape.
[9] https://www.awaresystems.be/imaging/tiff/bigtiff.html.
[10] https://openslide.org/.

Table 4
Evaluation on combinations of different strategies. The symbol ✓ denotes taking one strategy while symbol ✗ means taking the converse strategy. The best results are highlighted in bold.

Feature	Boxes	Sample	glo-det	art-det	tub-det	all-det	glo-ins	art-ins	tub-ins	all-ins
✗	✗	✗	–	–	–	–	–	–	–	–
✗	✗	✓	–	–	–	–	–	–	–	–
✗	✓	✗	75.8	24.9	56.4	52.4	–	–	33.6	–
✗	✓	✓	76.1	**26.2**	57.6	53.3	7.3	4.8	35.2	15.8
✓	✗	✗	–	–	–	–	–	–	–	–
✓	✗	✓	8.5	4.3	13.5	8.8	6.8	4.1	12.5	7.8
✓	✓	✗	12.6	6.8	56.7	25.4	10.4	5.3	45.8	20.5
✓	✓	✓	**80.9**	25.7	**58.5**	**55.1**	**74.3**	**23.4**	**55.8**	**51.2**

glo: glomeruli. *art*: arteries. *tub*: tubuli. *all*: overall. *det*: detection. *ins*: instance. "–": not working.
feature: if use separate feature maps between bbox-decoder and mask-decoder.
boxes: if use ground-truth boxes in the sampling process.
sample: if have class balance in the sampling process.

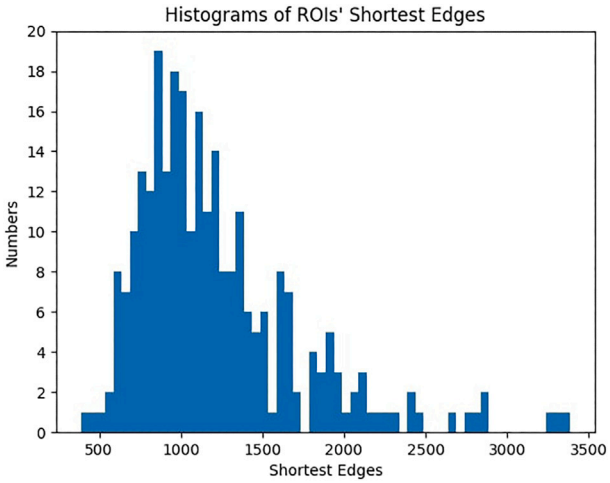


Fig. 8. The histogram of the shortest edges of ROIs extracted from WSIs. In our dataset, the shortest ROI length is 387 pixels, and the largest is 4565 pixels.

Table 5
Evaluation of detection performances on different types of queries. ∗ best result of each type is highlighted in bold.

mAP \ Number	200	300	400	500	600
Dynamic	**33.6**	**39.4**	**45.8**	**52.1**	**51.7**
Static	23.8	38.9	37.4	34.6	31.6

Table 6
Lesion dataset configuration and classification Performance.

	Tubuli		Glomeruli	
	Normal	Atrophy	Normal	Sclerosis
Train	10 326	10 647	534	382
Eval	906	568	51	22
per_precision	85.1		93.2	
per_recall	40.5		88.7	
all_precision	89.2			
all_recall	64.6			

shortest edge between 480 and 1333 pixels, while *RandomCrop* extracted patches with the shortest edge between 384 and 600 pixels. Finally, pixel values were normalized to $(-1, 1)$. As **mmdetection** supports training with variable image sizes, all models were allowed to train on objects at multiple scales.

For feature extraction, we used 256 channels for queries. During both training and inference, DiffRegFormer employed **500** dynamic queries, while QueryInst used 300 and 500 static queries, respectively. Cascade Mask-RCNN, QueryInst, and our method utilized 6 cascading stages. Table 1 presents performance comparisons for object detection

Table 7
Overall metrics for method assessment. *T* stands for trillion floating-point operations per second. *s* means second in time. *Para* refers to the number of parameters where *M* denotes million.

		MASD	DICE	FLOPS (T)	Speed (s)	Para (M)
One-stage	QueryInst_300	7.49	42.6	0.41	1.425	173
	QueryInst_500	10.95	35.9	0.55	1.647	246
Two-stage	Mask-RCNN	4.78	50.7	0.20	**0.864**	**43.98**
	CMask-RCNN	3.87	52.1	**1.72**	1.156	77.03
Our model	DiffRegFormer	**3.64**	**56.6**	0.55	1.324	140

w.r.t bounding boxes and instance segmentation w.r.t masks, respectively. For all experiments, DiffRegFormer did not utilize refinement during inference (i.e., iteration set to 1). Notably, Mask-RCNN [43] and cascade Mask-RCNN [62] are two-stage methods employing RPN networks instead of dynamic queries. QueryInst is a one-stage method with static queries. In contrast, DiffRegFormer is categorized as a one-stage method with dynamic queries. We adopted the Average Precision (AP) used in MSCOCO [30] as the metric to evaluate the models' performance in detection and segmentation.

Detection. DiffRegFormer achieves an Average Precision (AP) of 52.1% for detection with a ResNet-50 backbone, thereby surpassing competitors such as QueryInst, Mask R-CNN, and Cascade Mask R-CNN by significant margins. We observe that the iterative refinement hinders detection performance for small objects while benefiting the detection of medium to large objects. It suggests additional operations to mitigate information loss for small objects during the refinement process. Furthermore, static queries are less effective for handling dense objects, particularly when the number of queries significantly exceeds the number of feature channels (300 or 500 compared to 256), leading to a marked decline in detection performance.

Instance segmentation. DiffRegFormer achieves a 46.8% AP for instance segmentation and thus outperforms its counterparts. Moreover, applying iterative refinement to bounding boxes significantly impacts the instance segmentation outcomes. This effect can be attributed to the binary segmentation process within the boxes, where the quality of the instance masks is inherently linked to the recall and precision rates of the bounding boxes.

Table 2 presents the performance per object class. Our model demonstrates an overall superiority in detecting all anatomical structures compared to other methods. Additionally, it performs better in the instance segmentation of glomeruli and arteries while keeping a slight advantage in tubuli segmentation. These findings indicate the ability of our model to process dense, multi-class objects with diverse scales at the ROI level. However, for each model, we observe a performance bottleneck in the processing of arteries, posing a limitation for clinical application due to the significant size variation within the dataset. For instance, the largest arteries are two orders of magnitude larger than the smallest. This fact suggests that more sophisticated attention

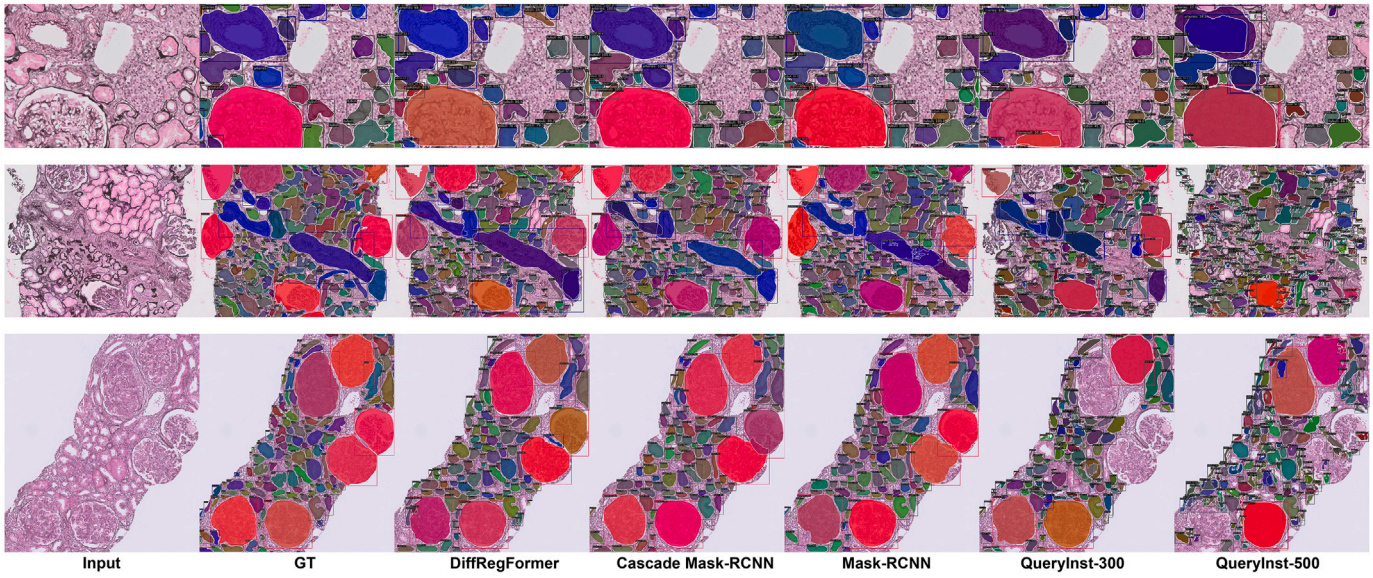


Fig. 9. A visual comparison of different instance segment methods on Jones-stained kidney biopsies. It illustrates that our model, i.e., DiffRegFormer, outperforms in detecting objects of diverse size and shape. Despite some instances overlapping, it generates precise instance masks over each region. Mask-RCNN-based models struggle to localize one large object within a single bounding box as they lack attention mechanisms to capture long-range dependencies. QueryInst with static queries fails to process dense objects since it fails to detect objects of each class within one ROI.

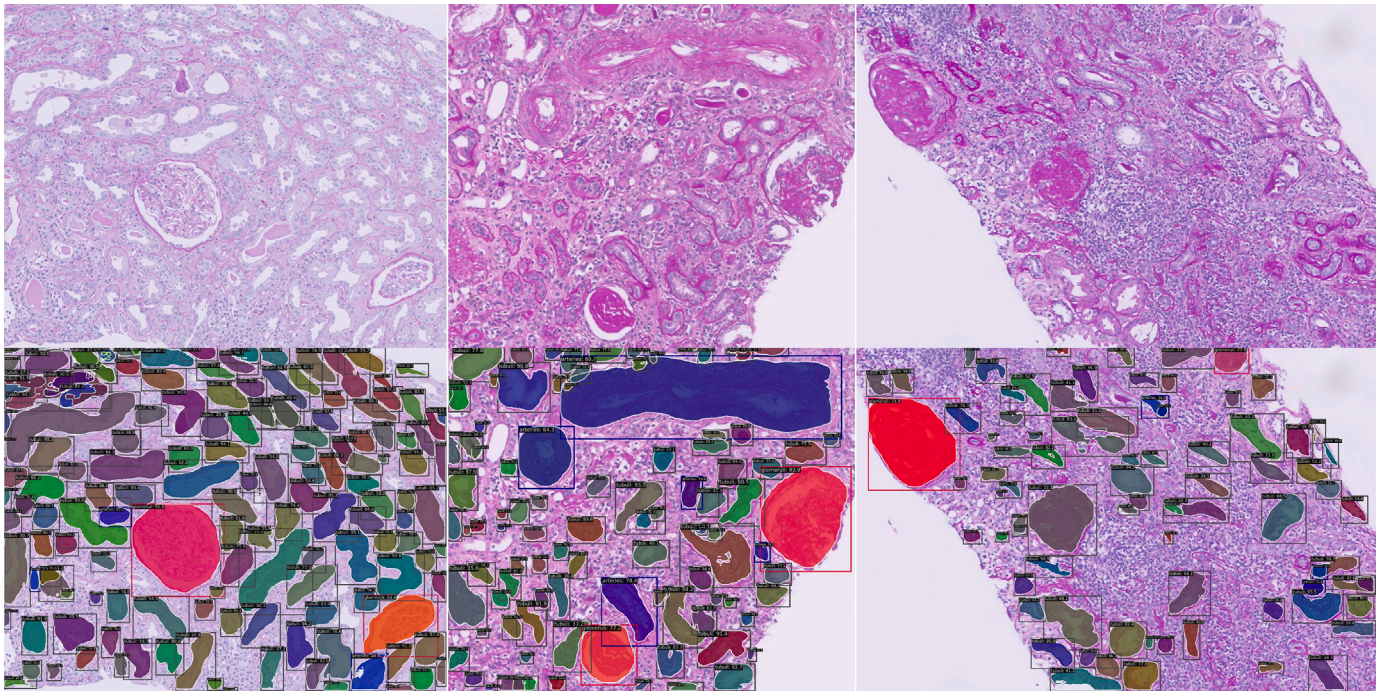


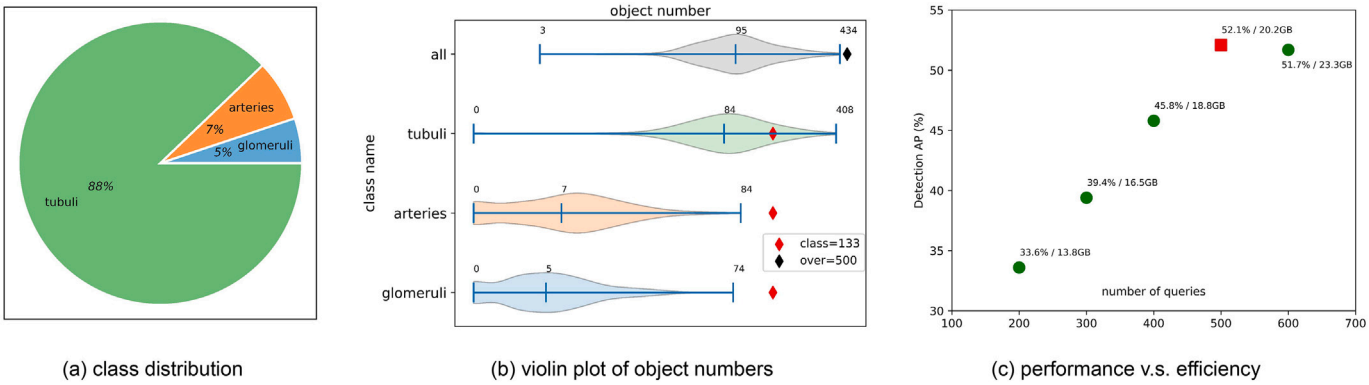
Fig. 10. A visualization of results of our model (DiffRegFormer) on the PAS-stained kidney biopsies with increasing degrees of fibrosis and inflammation (from left to right), along with variations in stain intensity.

Table 8
Overall evaluation result. The semantic method is an instance segmentation framework that combines semantic segmentation with a post-processing operation. We have reproduced it following the literature [18]. *AP* denotes average precision. AP_{50} and AP_{75} denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively.

	Bounding boxes			Instance			Extra metrics				
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	DICE	MASD	FLOPS (T)	Speed (s)	Para (M)
DiffRegFormer	**52.1**	**71.1**	**57.7**	**46.8**	**71.6**	**52.8**	**56.6**	**3.87**	0.55	1.324	140
Semantic method	44.6	64.8	46.7	40.3	65.3	44.6	45.4	7.86	**1.24**	**0.760**	**32.8**

mechanisms are needed to capture global contextual information for these extreme variations in scale.

Fig. 9 demonstrates DiffRegFormer's capability to detect objects at various scales effectively. It accurately generates instance masks



(a) class distribution (b) violin plot of object numbers (c) performance v.s. efficiency

Fig. 11. Complexity Analysis on Jones-stained renal biopsy dataset. Figure (a) is the sample occupation across our dataset. Figure (b) is a violin plot showing the sample number distribution per class. Figure (c) shows trade-offs between performance and efficiency for different experiment settings. The red square is the optimal setting, while the green circles are the sub-optimal settings.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Evaluation results per object class. *AP* denotes average precision. AP_{50} and AP_{75} denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively.

	Bounding boxes - glo			Instance - glo			Extra metrics	
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}	DICE	MASD
DiffRegFormer	80.9	94.2	89.8	74.3	94.2	80.9	78.8	1.21
semantic method	75.6	82.4	81.2	68.3	84.3	70.1	70.6	8.68
	Bounding boxes - art			Instance - art			Extra metrics	
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	DICE	MASD
DiffRegFormer	25.7	46.9	25.8	23.4	48.8	20.2	30.2	25.94
semantic method	10.3	25.8	13.6	11.5	24.9	10.3	16.3	49.14
	Bounding boxes - tub			Instance - tub			Extra metrics	
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	DICE	MASD
DiffRegFormer	58.5	80.2	67.9	55.8	80.6	62.3	64.8	0.89
semantic method	43.6	73.5	54.6	46.9	75.2	52.4	50.5	2.16

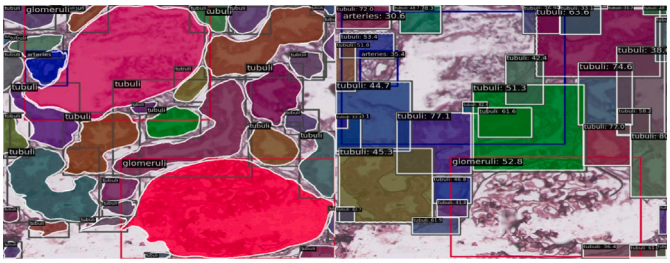
within bounding boxes, demonstrating precise localization of larger arteries (see the second row). In contrast, Cascade Mask R-CNN has difficulty locating larger objects within a single bounding box due to the lack of an attention mechanism for capturing long-range dependencies. Furthermore, Mask-RCNN tends to generate multiple detection results for a single object(see the third row) due to the absence of cascade refinement. These observations suggest that iterative bounding box refinement during training significantly reduces multiple predictions for a single object. At the same time, the attention mechanism with dynamic queries processes large-scale objects. Nevertheless, QueryInst, with its reliance on static queries, may struggle to effectively detect all instances in datasets with significant size and shape variations. QueryInst-300 tends to miss many structures, while QueryInst-500 can even degenerate into generating multiple false predictions for a single object. Furthermore, RCNN-based models share a limitation: a single pixel may be allocated to multiple instance masks due to pixel-wise predictions within each independent ROI. It inherently prevents enforcing the spatial exclusivity of pixels to instances. Additionally, Fig. 10 depicts DiffRegFormer's potential for domain transfer, as it generates reasonable results on PAS-stained images while being trained exclusively on Jones' silver-stained images. Moreover, Fig. 14 highlights that DiffRegFormer, combined with sliding window and stitching techniques, can be effectively applied to WSIs, further demonstrating its effectiveness.

Lesion classification. The data augmentation pipeline for training the lesion classifier is relatively straightforward. We set the batch size as 96 for the input croppings, where 32 patches for each class. Each cropping first undergoes *RandomFlip* with a probability of 0.5, followed by resizing to 256 × 256 pixels. Finally, pixel values are normalized to the $(-1, 1)$ range. As shown in Table 6, the classifier achieves an overall precision of 89.2% and a recall of 64.6%. More specifically, per-class precision for atrophic tubuli is 85.1%, and 93.2% for sclerotic glomeruli. In addition, per-class recall is 40.5% for atrophic tubuli and 88.7% for sclerotic glomeruli. Since the lesion classifier operates on croppings of the anatomical structure generated by DiffRegFormer during inference, its recall rate heavily depends on the performance of our dense instance segmentation model. However, the high precision indicates that lesion classification on correctly cropped structures is very accurate. In summary, the enhancement of the performance of DiffRegFormer is expected to improve the recall of the lesion classifier further. Fig. 15 visualizes the mapping of lesion predictions onto the instance segmentation results. It reveals that our model can be a solid foundation for further clinical diagnosis.

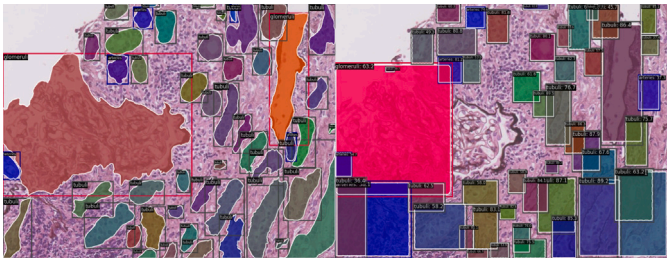
4.2.1. Metrics for instance segmentation

We have computed a set of commonly used metrics to assess the performance of our method in a broader scope. These are DICE, MASD (mean average surface distance), FLOPS, inference speed, and amount of parameters. In Table 7, we can observe from the values of DICE and MASD that our method outperforms other approaches on segmentation performance. The resource efficiency is, however, moderate. The main cause of resource cost is the attention mechanism, as seen in QueryInst and DiffRegFormer. To that end, it is beneficial to adopt more efficient attention algorithms to accelerate the inference speed and lessen resource burden, tailoring for the practical clinical agenda. In addition, the arteries that vary across multiple scales make it difficult to obtain high-quality instance masks. All methods have illustrated a large average surface distance between the predicted and ground-truth boundaries.

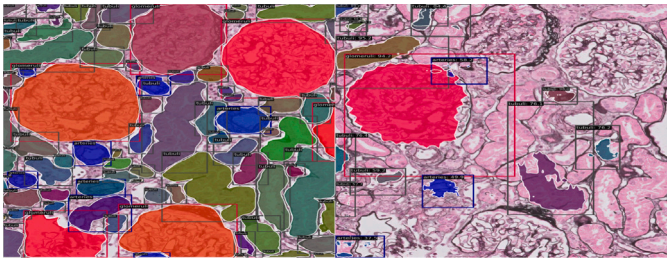
(a) shared feature + proposals + biased sampling on MSCOCO's image



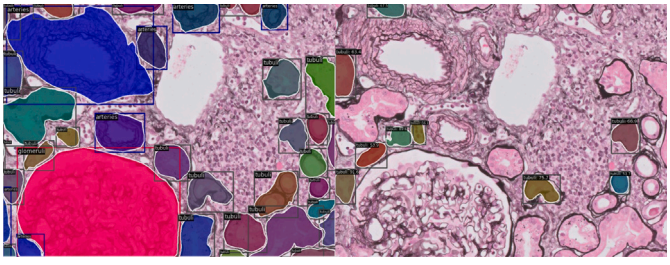
(b) shared feature + ground-truth + biased sampling on a ROI



(c) shared feature + ground-truth + unbiased sampling on a ROI



(d) separate feature + proposals + unbiased sampling on a ROI



(e) separate feature + ground-truth + biased sampling on a ROI

Fig. 12. Ablation study on combinations of different strategies. The images on the left are ground-truth. The images on the right are our predictions.

4.2.2. Comparison with segmentation based method

In order to supplement our analysis in terms of completeness, we have done a retrospective analysis reproducing a semantic segmentation based instance segmentation framework [18]. We denote this method as a semantic method and compare its results with ours. In Table 8, an overall comparison is depicted. Our method outperforms the semantic method in prediction quality at the cost of more inference time and computational resources. It should be noted these results are specific to this dataset. In Table 9, the superiority of our method per class is depicted. Furthermore, Fig. 16 illustrates that the semantic method is more vulnerable to the segmentation quality of the auxiliary boundary and interstitium class, which is insufficient in separating objects tightly touching. Such touching objects are typical in kidney tissue. Our method, on the contrary, can detect and segment each object.

4.3. Complexity assessment of model training

We start with analyzing the intrinsic class imbalance within our dataset. Of the total structures (21,889 objects), 88% are tubuli, followed by 7% arteries (1679 objects), and 5% glomeruli (1226 objects). This distribution is depicted in Fig. 11a, highlighting the necessity of sampling to include more rare instances, such as arteries and glomeruli, among the dominant tubuli. A violin plot, shown in Fig. 11b, details the distribution of structure counts at the ROI level. On average, ROIs contain 95 objects, with a maximum of 434. To accommodate all objects, DiffRegFormer is configured to handle a maximum number exceeding 434 per image. Fig. 11c depicts the correlation between the increase in computational cost and performance improvement. As expected, we observe a turning point at 500 dynamic queries. This configuration yields a favorable trade-off between complexity and efficiency, achieving 52.1% AP with 20.2 GB GPU memory usage. Specifically, ROIs contain an average of 5 glomeruli, 7 arteries, and 84 tubuli with corresponding maximums of 74, 84, and 408. Setting the maximum number per class to 166 is another feasible option for positive sample generation, as our sampling approach ensures comprehensive inclusion of rare instances while maintaining the rational proportion of the majority class. In conclusion, selecting a maximum of 500 objects per image and 166 per class can effectively balance performance and computational efficiency.

4.4. Ablation study

The ablation study focuses on the effects of various training strategies and query types on model performance. By analyzing our results, we can gain insights into the specific characteristics of each strategy. This analysis highlights the significance of our improvements as key contributions to adapting an end-to-end instance segmentation model to datasets with dense objects.

4.4.1. Training strategies

Training an instance segmentation model for datasets with dense objects, such as, in our case, kidney biopsies, presents distinct challenges compared to the common datasets with sparser object distributions like MSCOCO. Instance segmentation models typically employ three strategies: (1) shared features between the bbox-decoder and mask-decoder; (2) selection of positive samples from proposal bounding boxes; and (3) class-imbalanced sampling. However, for dense object segmentation, we propose three alternative strategies: (a) separate features for the bbox and mask decoders; (b) positive sample selection directly from ground-truth boxes; and (c) class-balanced sampling. While the conventional combination of shared features, proposal boxes, and biased sampling is adequate for standard instance segmentation (see Fig. 12(a)), only the combination of separate features, ground-truth boxes, and unbiased sampling proves to be successful for dense instance segmentation. Table 4 evaluates the combinations above, with conventional strategies marked with ✗ and our proposed strategy with ✓. Notably, only five out of eight possible combinations can train DiffRegFormer.

Fig. 12(b) demonstrates that combining shared features, ground-truth boxes, and biased sampling enables object detection across all classes. However, it fails to produce accurate instance masks within

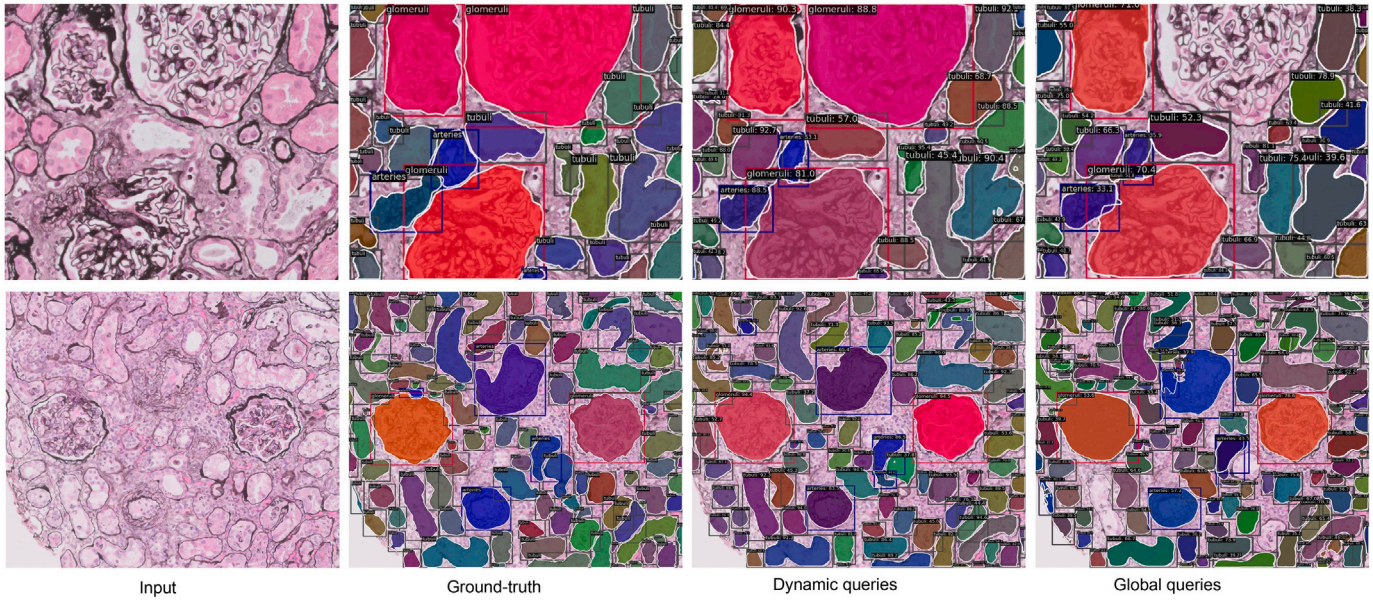


Fig. 13. A visual comparison from an ablation study on different query types. Dynamic queries outperform static queries on detection. It should be noted that the static queries fail to detect objects of all classes within one ROI.

these bounding boxes. The observation suggests that the bbox-decoder dominates the shared feature representation, thus hindering the mask decoder's ability to extract pixel-wise information. Consequently, all pixels within detected boxes are classified as tubuli. Moreover, biased sampling impedes the mask decoder's capacity to learn features for rare instances (glomeruli and arteries), resulting in no pixel classification within those boxes.

Additionally, Fig. 12(c), shared features combined with ground-truth boxes and unbiased sampling can detect and segment objects across all classes, albeit the degraded mask quality. This degradation is likely due to the bbox-decoder's dominance within the shared feature representation.

In contrast, Fig. 12(d) reveals that using separate features, proposal boxes, and unbiased sampling leads to inaccurate object detection. It is probably due to the accumulation of errors introduced by proposals during the early stages of diffusion model training.

Lastly, Fig. 12(e) demonstrates that separate features, ground-truth boxes, and biased sampling result in correct detection and segmentation for only a small proportion of tubuli. This finding further highlights the challenges caused by biased sampling in learning features for rare instances.

In summary, shared feature maps lead to the bbox-decoder dominating feature representation learning, impairing the mask-decoder's ability to extract pixel-wise information necessary for accurate binary instance masks. Biased sampling further hinders the mask-decoder from learning information to segment rare instances (glomeruli and arteries), causing it to be overwhelmed by the majority class (tubuli). Finally, utilizing proposals for sampling instead of ground-truth boxes introduces cumulative errors that disrupt the training for both decoders, preventing the learning of contextual information necessary for long-range correlations via attention mechanisms.

4.4.2. Queries

Static queries in transformer-based models face challenges in representing dense objects within large-scale datasets due to their static nature. Adequate accommodation to extensive variations in object size and shape necessitates a proportional increase in query count. Alternatively, dynamic queries, generated on-the-fly for each object in a mini-batch, demonstrate superior adaptability to dataset complexity. Theoretically, assigning more static queries could achieve comparable performance to dynamic queries. However, our experiments indicate that increasing the number of static queries without a corresponding increase in feature channels does not yield improvements. Consequently, employing static queries for dense object processing in transformer-based models is significantly less cost-effective than dynamic queries.

Table 5 delineates our comparative analysis of varying query counts for both dynamic and static queries, with the number of feature channels fixed at 256. We observe that increasing the number of dynamic queries leads to a gain in performance while increasing the number of static queries results in performance degradation. Fig. 13 further illustrates the inherent challenge of achieving a favorable trade-off between complexity and performance with static queries. Even with 500 queries, the model with static queries fails to detect structures across all classes consistently. This limitation is not attributable to class imbalance compared to the dynamic query strategy. In conclusion, the dynamic query is a more efficient approach for datasets with dense objects, highlighting their superiority over static queries.

5. Conclusion and future work

Our research introduces a novel framework to process large-scale renal WSI datasets with potential changes in lesion combinations. Our model performs better in predicting lesions within dense objects, effectively handling multi-class and multi-scale challenges at the ROI level. We present the first dense instance segmentation module, DiffRegFormer, that seamlessly integrates a diffusion model with a transformer in RCNN-style. This approach leverages the diffusion model to generate refined object boxes, upon which a regional transformer creates dynamic queries that are ultimately transformed into accurate instance masks within the boxes. Furthermore, our lesion classifier accurately predicts class-specific lesions for each cropped structure. The model demonstrates several advantages in handling dense object datasets: (1) its iterative noise-to-box approach eliminates the dependence on prior knowledge about object size and shape, facilitating rapid object localization; (2) dynamic queries, generated on-the-fly for each object in a mini-batch style, ensuring scalability to large datasets; (3) the modular design of the model, with independent plug-in prediction heads, allows for replacement and promotes adaptability. This flexibility enables the efficient reuse of trained components when up-scaling to large-scale datasets. Experimental results on Jones' stained renal biopsies

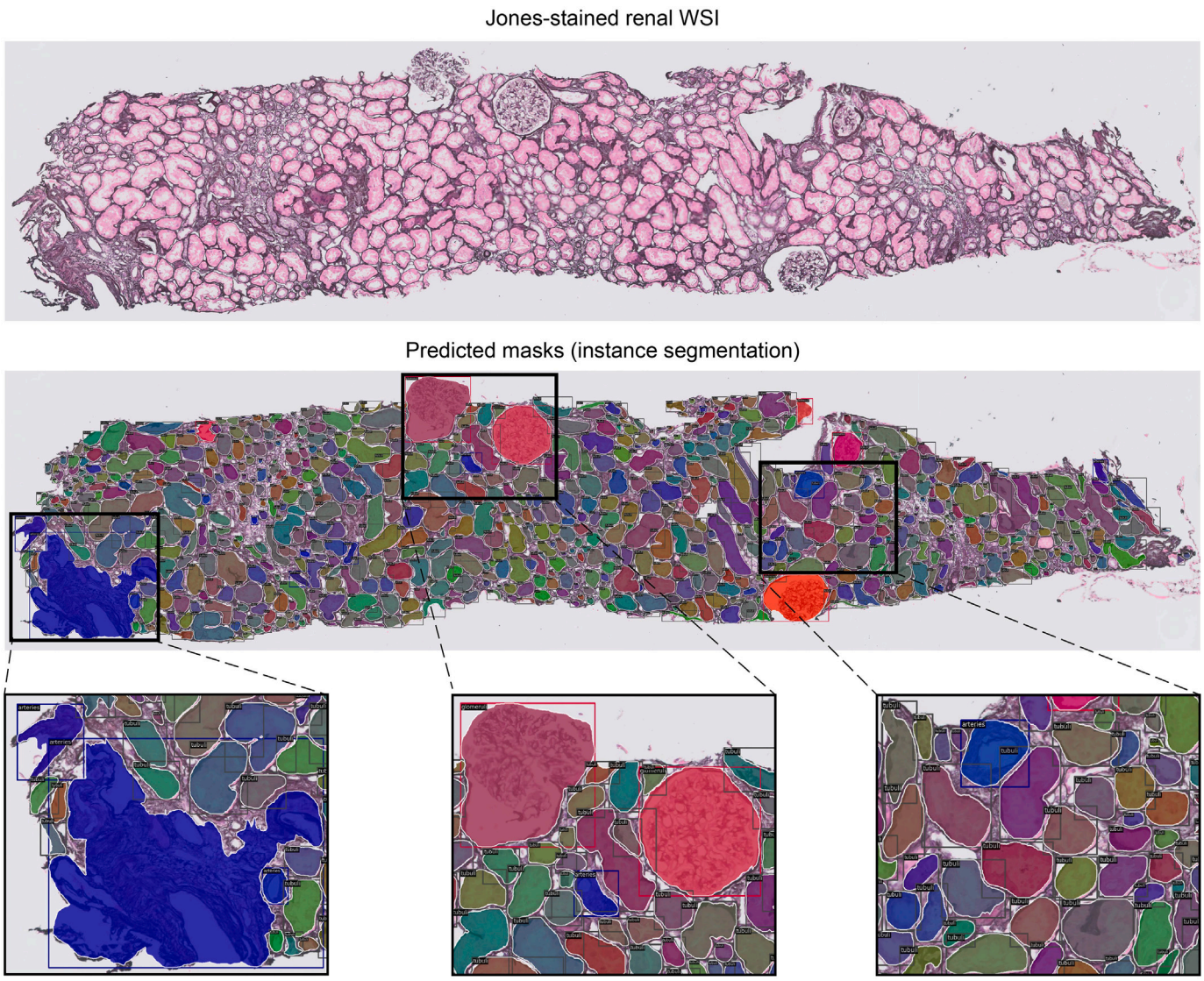


Fig. 14. A representative visualization of results from DiffRegFormer stitches showing predicted dense instance objects from a WSI of a Jones-stained renal biopsy.

demonstrate that our model surpasses existing benchmarks. Comparing results between methods on the basis of the data is tricky. Measures like DICE behave differently with different datasets. Comparison should be on the same dataset (cf. 4.2.2). In that case, results can be well understood and objectively compared.

The current model needs further exploration to meet the rigorous requirements for clinical applications. Still, it establishes a robust framework for dense, multi-class, multi-scale object recognition at the ROI level, setting a solid foundation for future computational improvements. Potential areas for future research include: (1) further enhancing the compatibility of diffusion models with transformers like DETR to avoid NMS post-processing. (2) developing innovative regional feature extractors to generate more informative dynamic queries and mitigate the loss of information for small objects during iterative refinements. (3) integrating spatial constraints into regional mask inference to prevent erroneous assignment of single pixels to multiple instances. (4) designing lightweight attention mechanisms to lessen the computational burden. In conclusion, the approach presented paves the way for efficient and reliable automated assessment as a tool for the workbench of renal pathologists.

CRediT authorship contribution statement

Zhan Xiong: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Junling He:** Writing – review & editing, Validation, Resources, Data curation. **Pieter Valkema:** Software, Resources, Data curation. **Tri Q. Nguyen:** Resources, Data curation. **Maarten Naesens:** Resources, Data curation. **Jesper Kers:** Writing – review & editing, Resources, Funding acquisition, Data curation. **Fons J. Verbeek:** Writing – review & editing, Validation, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

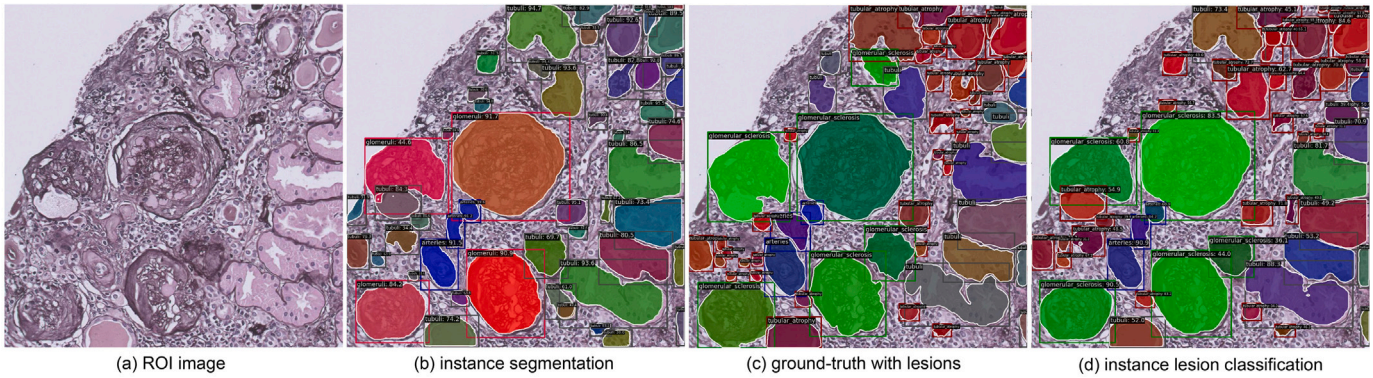


Fig. 15. A visualization of results of the lesion classifier predicting the probability of class-wise lesions on an ROI. Combined with dense instance segmentation, our model can generate per-instance lesion identification.

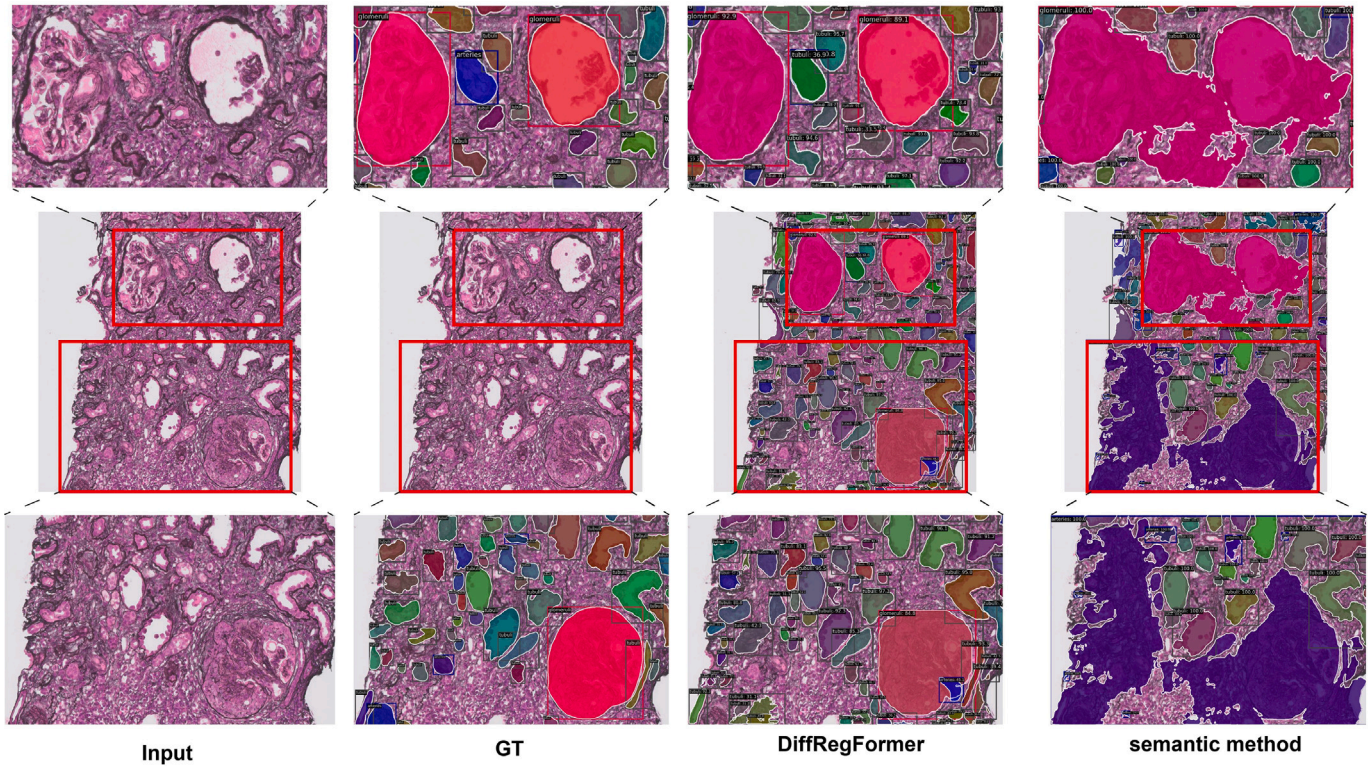


Fig. 16. A visual comparison between DiffRegFormer and the semantic method. The upper row represents a zoom of the upper-box in the middle row. The lower row represents a zoom of the lower-box in the middle row. The quality of the result w.r.t. the ground truth labels can be convincingly assessed in favor of DiffRegFormer.

Acknowledgments

This work is supported in part by funds from the Dutch Kidney Foundation, Netherlands (17OKG23), the Research Priority Area Human(e) AI at the University of Amsterdam, Netherlands, and the Chinese Scholarship Council (CSC).

Appendix. Additional visualization samples

As an extra, we have added two zoom-in samples to highlight the details of our method and other approaches. The first row of Figure A.17 shows that our method can correctly divide two glomeruli without overlap. Cascade Mask RCNN and Mask RCNN, however, generate multiple responses at the border of two objects. QueryInst with static queries cannot detect and segment any glomerulus. The last row of Figure A.17 demonstrates four large arteries in a line. It is complicated since two large arteries pass over most input image regions. Although it had slightly multiple responses at the artery border, our method can capture long-range spatial dependencies for large objects using attention mechanisms with dynamic queries. In contrast, Cascade Mask RCNN and Mask RCNN cannot model large objects without attention modules. QueryInst adopts an attention mechanism with static queries and cannot process complicated combinations of shapes and appearances with fixed query numbers (see Fig. A.17).

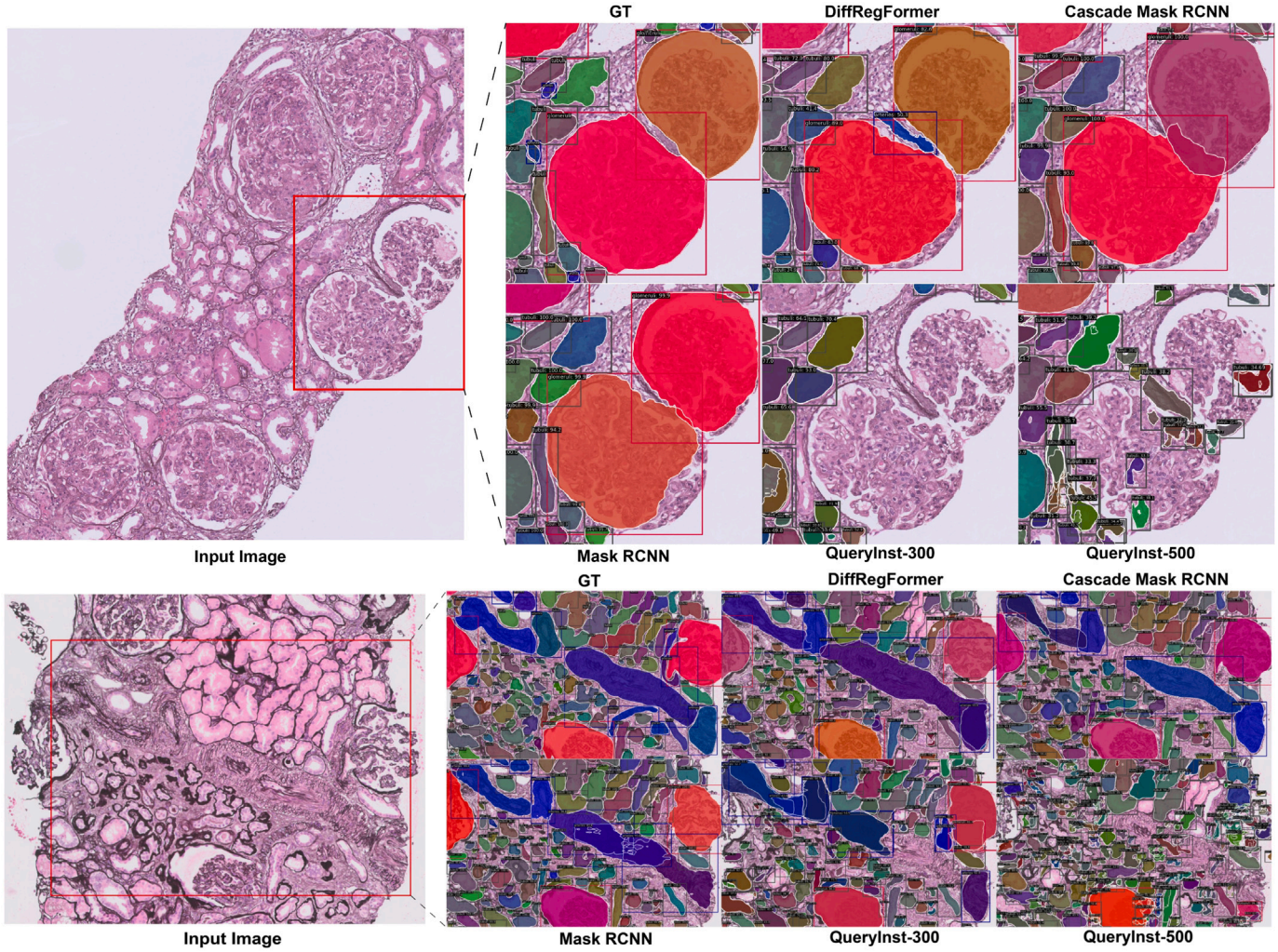


Fig. A.17. Detailed zoom-in samples comparing DiffRegFormer with other methods.

References

[1] Brachemi S, Bollée G. Renal biopsy practice: What is the gold standard? World J Nephrol 2014;3(4):287.

[2] Alnazer I, Bourdon P, Urruty T, Falou O, Khalil M, Shahin A, Fernandez-Maloigne C. Recent advances in medical image processing for the evaluation of chronic kidney disease. Med Image Anal 2021;69:101960.

[3] Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: A survey. Med Image Anal 2021;67:101813.

[4] Xu Y, Mo T, Feng Q, Zhong P, Lai M, Eric I, Chang C. Deep learning of feature representation with multiple instance learning for medical image analysis. In: 2014 IEEE international conference on acoustics, speech and signal processing. ICASSP, IEEE; 2014, p. 1626–30.

[5] Bouteldja N, Klinkhammer BM, Bülow RD, Droste P, Otten SW, von Stillfried SF, Moellmann J, Sheehan SM, Korstanje R, Menzel S, et al. Deep learning–based segmentation and quantification in experimental kidney histopathology. J Am Soc Nephrol.: JASN 2021;32(1):52.

[6] Jiang L, Chen W, Dong B, Mei K, Zhu C, Liu J, Cai M, Yan Y, Wang G, Zuo L, et al. A deep learning-based approach for glomeruli instance segmentation from multistained renal biopsy pathologic images. Am J Pathol 2021;191(8):1431–41.

[7] Salvi M, Mogetta A, Gambella A, Molinaro L, Barreca A, Papotti M, Molinari F. Automated assessment of glomerulosclerosis and tubular atrophy using deep learning. Comput Med Imaging Graph 2021;90:101930.

[8] Deng R, Liu Q, Cui C, Yao T, Long J, Asad Z, Womick RM, Zhu Z, Fogo AB, Zhao S, et al. Omni-seg: A scale-aware dynamic network for renal pathological image segmentation. IEEE Trans Biomed Eng 2023.

[9] Yuan M, Xia Y, Dong H, Chen Z, Yao J, Qiu M, Yan K, Yin X, Shi Y, Chen X, et al. Devil is in the queries: advancing mask transformers for real-world medical image segmentation and out-of-distribution localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 23879–89.

[10] Lin G, Zhang Z, Long K, Zhang Y, Lu Y, Geng J, Zhou Z, Feng Q, Lu L, Cao L. GCLR: A self-supervised representation learning pretext task for glomerular filtration barrier segmentation in TEM images. Artif Intell Med 2023;146:102720.

[11] Meseguer P, Del Amor R, Naranjo V. MICIL: Multiple-instance class-incremental learning for skin cancer whole slide images. Artif Intell Med 2024;152:102870.

[12] Feng C, Ong K, Young DM, Chen B, Li L, Huo X, Lu H, Gu W, Liu F, Tang H, et al. Artificial intelligence-assisted quantification and assessment of whole slide images for pediatric kidney disease diagnosis. Bioinform. 2024;40(1):btad740.

[13] Gadermayr M, Gupta L, Appel V, Boor P, Klinkhammer BM, Merhof D. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. IEEE Trans Med Imaging 2019;38(10):2293–302.

[14] Vasiljević J, Feuerhake F, Wemmert C, Lampert T. Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks. Neurocomputing 2021;460:277–91.

[15] Kang H, Luo D, Feng W, Zeng S, Quan T, Hu J, Liu X. Stainnet: a fast and robust stain normalization network. Front Med 2021;8:746307.

[16] Zhang Y, Li X, Chen H, Yuille AL, Liu Y, Zhou Z. Continual learning for abdominal multi-organ and tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023, p. 35–45.

[17] Deng R, Liu Q, Cui C, Yao T, Yue J, Xiong J, Yu L, Wu Y, Yin M, Wang Y, et al. PrPSeg: Universal proposition learning for panoramic renal pathology segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 11736–46.

[18] Hermsen M, de Bel T, Den Boer M, Steenbergen EJ, Kers J, Florquin S, Roelofs JJ, Stegall MD, Alexander MP, Smith BH, et al. Deep learning–based histopathologic assessment of kidney tissue. J Am Soc Nephrol.: JASN 2019;30(10):1968.

[19] Liu X, Wu Y, Chen Y, Hui D, Zhang J, Hao F, Lu Y, Cheng H, Zeng Y, Han W, et al. Diagnosis of diabetic kidney disease in whole slide images via AI-driven quantification of pathological indicators. Comput Biol Med 2023;166:107470.

[20] Jha A, Yang H, Deng R, Kapp ME, Fogo AB, Huo Y. Instance segmentation for whole slide imaging: end-to-end or detect-then-segment. J Med Imaging 2021;8(1). 014001–014001.
[21] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European conference on computer vision. Springer; 2020, p. 213–29.
[22] Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 1290–9.
[23] Shickel B, Lucarelli N, Rao A, Yun D, Moon KC, Seok HS, Sarder P. Spatially aware transformer networks for contextual prediction of diabetic nephropathy progression from whole slide images. In: Medical imaging 2023: digital and computational pathology. Vol. 12471, SPIE; 2023, p. 129–40.
[24] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 2020;33:6840–51.
[25] Chen C-F, Panda R, Fan Q. Regionvit: Regional-to-local attention for vision transformers. 2021, arXiv preprint arXiv:2106.02689.
[26] Li F, Zhang H, Liu S, Guo J, Ni LM, Zhang L. Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 13619–27.
[27] Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, Ni LM, Shum H-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. 2022, arXiv preprint arXiv:2203.03605.
[28] Cheng T, Wang X, Chen S, Zhang W, Zhang Q, Huang C, Zhang Z, Liu W. Sparse instance activation for real-time instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 4433–42.
[29] Li F, Zhang H, Xu H, Liu S, Zhang L, Ni LM, Shum H-Y. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 3041–50.
[30] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. Springer; 2014, p. 740–55.
[31] Gonzalez RC, Woods RE. Digital Image Processing (3rd Edition). USA: Prentice-Hall, Inc.; 2006.
[32] Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models. In: International conference on machine learning. PMLR; 2021, p. 8162–71.
[33] Huang Q, Zambrini J-C. Stochastic geometric mechanics in nonequilibrium thermodynamics: Schrödinger meets onsager. J Phys A 2023;56(13):134003.
[34] Oh H-J, Jeong W-K. Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023, p. 337–45.
[35] Li J, Zheng S, Zhu C, Sun Y, Chen P, Shui Z, Zhang Y, Li H, Yang L. PathUp: Patch-wise timestep tracking for multi-class large pathology image synthesising diffusion model. In: Proceedings of the 32nd ACM international conference on multimedia. 2024, p. 3984–93.
[36] Li X, Thickstun J, Gulrajani I, Liang PS, Hashimoto TB. Diffusion-lm improves controllable text generation. Adv Neural Inf Process Syst 2022;35:4328–43.
[37] Popov V, Vovk I, Gogoryan V, Sadekova T, Kudinov M. Grad-TTS: A diffusion probabilistic model for text-to-speech. 2021, arXiv:2105.06337.
[38] Fan J, Lv T, Wang P, Hong X, Liu Y, Jiang C, Ni J, Li L, Pan X. DCDiff: Dual-granularity cooperative diffusion models for pathology image analysis. IEEE Trans Med Imaging 2024.
[39] Chen S, Sun P, Song Y, Luo P. Diffusiondet: Diffusion model for object detection. 2022, arXiv preprint arXiv:2211.09788.
[40] Gu Z, Chen H, Xu Z, Lan J, Meng C, Wang W. Diffusioninst: Diffusion model for instance segmentation. 2022, arXiv preprint arXiv:2212.02773.
[41] Girshick R. Fast R-CNN. 2015, arXiv:1504.08083.
[42] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. 2016, arXiv:1506.01497.
[43] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 2961–9.
[44] Huang Z, Huang L, Gong Y, Huang C, Wang X. Mask scoring R-CNN. 2019, arXiv:1903.00241.
[45] Fang Y, Yang S, Wang X, Li Y, Fang C, Shan Y, Feng B, Liu W. Instances as queries. 2021, arXiv:2105.01928.
[46] Song J, Meng C, Ermon S. Denoising diffusion implicit models. 2020, arXiv preprint arXiv:2010.02502.
[47] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 10684–95.
[48] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. PMLR; 2015, p. 2256–65.
[49] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015, arXiv:1512.03385.
[50] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009, p. 248–55.
[51] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. 2017, arXiv:1612.03144.
[52] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[53] Chen Y, Dai X, Liu M, Chen D, Yuan L, Liu Z. Dynamic convolution: Attention over convolution kernels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 11030–9.
[54] Cheng B, Schwing A, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. Adv Neural Inf Process Syst 2021;34:17864–75.
[55] Chen T, Li L, Saxena S, Hinton G, Fleet DJ. A generalist framework for panoptic segmentation of images and videos. 2022, arXiv preprint arXiv:2210.06366.
[56] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Adv Neural Inf Process Syst 2021;34:8780–94.
[57] Chen T, Zhang R, Hinton G. Analog bits: Generating discrete data using diffusion models with self-conditioning. 2022, arXiv preprint arXiv:2208.04202.
[58] Sun P, Jiang Y, Xie E, Shao W, Yuan Z, Wang C, Luo P. What makes for end-to-end object detection? In: International conference on machine learning. PMLR; 2021, p. 9934–44.
[59] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection. 2020, arXiv preprint arXiv:2010.04159.
[60] Mohamed S, Lakshminarayanan B. Learning in implicit generative models. 2016, arXiv preprint arXiv:1610.03483.
[61] Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, et al. MMDetection: Open mmlab detection toolbox and benchmark. 2019, arXiv preprint arXiv:1906.07155.
[62] Cai Z, Vasconcelos N. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans Pattern Anal Mach Intell 2019;43(5):1483–98.