Automatic fine-grained glomerular lesion recognition in kidney pathology

Yang Nan [a,1,*], Fengyi Li [b,1], Peng Tang [c], Guyue Zhang [b], Caihong Zeng [d], Guotong Xie [b,*], Zhihong Liu [d,*], Guang Yang [a,*]

[a] National Heart and Lung Institute, Imperial College London, London, United Kingdom
[b] PingAn Healthcare Technology, Shanghai, China
[c] Department of Informatics, Technical University of Munich, Munich, Germany
[d] National Clinical Research Center of Kidney Diseases, Nanjing University School of Medicine, Nanjing, China

ARTICLE INFO

ABSTRACT

Recognition of glomeruli lesions is the key for diagnosis and treatment planning in kidney pathology; however, the coexisting glomerular structures such as mesangial regions exacerbate the difficulties of this task. In this paper, we introduce a scheme to recognize fine-grained glomeruli lesions from whole slide images. First, a focal instance structural similarity loss is proposed to drive the model to locate all types of glomeruli precisely. Then an Uncertainty Aided Apportionment Network is designed to carry out the fine-grained visual classification without bounding-box annotations. This double branch-shaped structure extracts common features of the child class from the parent class and produces the uncertainty factor for reconstituting the training dataset. Results of slide-wise evaluation illustrate the effectiveness of the entire scheme, with an 8–22% improvement of the mean Average Precision compared with remarkable detection methods. The comprehensive results clearly demonstrate the effectiveness of the proposed method.

1. Introduction

Immunoglobulin A nephropathy (IgAN) is the leading cause of chronic kidney disease worldwide, especially in Asian regions, with nearly 40% of patients developing the end-stage renal disease within decades. Patients with this nephropathy have varied histological lesions, ranging from crescentic glomerulonephritis, mesangial proliferation to global and segmental sclerosis. The five main types of structure ( ) in IgAN (Fig. 1) includes Neg (tubule and arteriole), GS (Global Sclerosis), C (Crescent), SS (Segmental Sclerosis) and NoA (None of Above). However, due to the collapse and proliferation of capillary loops, some pathological changes in IgAN share a high visual similarity that even pathologists cannot achieve satisfactory agreement. A previous study found a low intra-class correlation coefficients of recognizing SS (0.66) and C (0.46) in IgAN, given by three to five pathologists [1]. Evidently, there is an urgent need to shift the balance of pathological changes' identification towards more objective and quantification.

As digital pathology evolves, biopsy tissues can be scanned as whole slide images (WSIs) through micro-scanners. Meanwhile, the remarkable success of deep convolutional neural networks also provide pathways to address the above intractable issues. With these efforts, researchers have studied computer-aided diagnosis for renal pathological feature recognition, including glomeruli location and lesion classification. Current location methods are mainly based on the combination of patch-wise predictions, including detection [2] and segmentation [3] of numerous glomeruli in large patches, or binary classification (glomerulus or non-glomerulus) in small patches [4]. However, most existing studies typically focus on classifying sclerosis and non-sclerosis and do not involve complicated lesion types, indicating the insufficient exploration of IgAN. To identify complex pathological changes in IgAN, several difficulties need to be addressed. Firstly, each glomerulus should be accurately localized. In addition, glomerular lesions with low intraclass variances should be precisely classified, even under an imbalanced data distribution. Unfortunately, most fine-grained classification tasks were performed on natural images [5,6] with relatively unitary background regions, which are contrary to the com-

---

* Corresponding authors.
E-mail addresses: y.nan20@imperial.ac.uk (Y. Nan), xieguotong@pingan.com.cn (G. Xie), liuzhihong@nju.edu.cn (Z. Liu), g.yang@imperial.ac.uk (G. Yang).
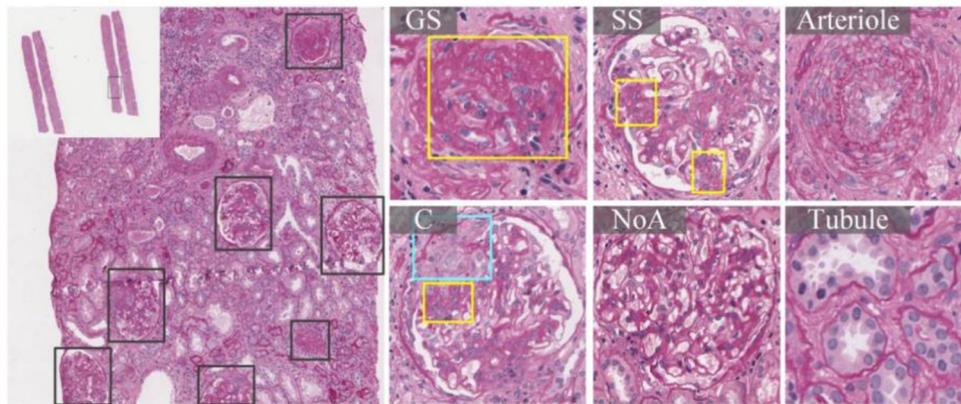[1] These authors contributed equally to this work.

Fig. 1. Glomeruli in IgAN whole slide images with periodic acid–Schiff stain. Pathological changes of sclerosis and crescent are marked with yellow and cyan boxes, respectively.
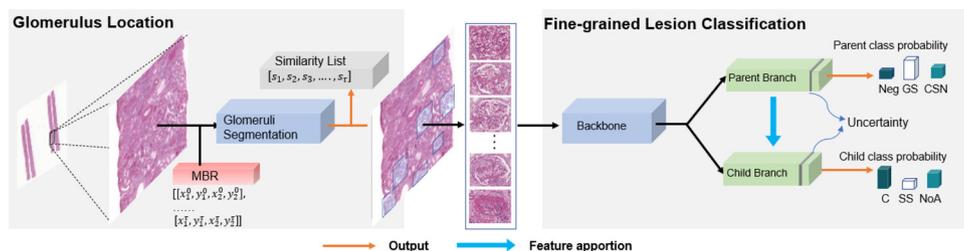

Fig. 2. Overview of the proposed scheme. The fine-grained detection task is divided into two sub-tasks, including glomerulus location and fine-grained classification.

plex background regions in renal pathological images (Supplement-Fig. 1). Besides, methods that heavily rely on part annotations [7], such as bounding boxes or masks that label the sub-regions, are impractical for medical images due to the high cost of manual annotations. For fine-grained recognition, the network should assess the 'confidence' of its prediction to express the certainty of its output. Unfortunately, methods such as Bayesian deep learning and Monte Carlo dropout require repetitious inferences, which are inconvenient and time-consuming. Last but not least, the network applied in medical image analysis should be able to reduce the negative impact of inconclusive annotations.

In this paper, we present the first attempt to conduct fine-grained visual recognition (FGVR) in large-scaled whole slide images, aiming to recognize complex pathological changes in IgAN. Different from previous works, FGVR aims to locate but distinguish fine-grained subcategories. To address the weak capacity of existing detection modules (such as Mask R-CNN [8], FCOS [9], etc.) on FGVR task, we propose a two-stage scheme, with glomerulus segmentation and classification, respectively. Initially, a focal instance structural similarity (FISS) loss is presented to acquire accurate segmentation results. It coalesces focal loss with instance structural similarity loss to accurately segment the boundary of the glomerulus. Then, the glomerular lesions are divided into two groups based on pathological representation and lesion severity, with Neg, GS, CSN (the combination of C, SS, and NoA) as a parent class and C, SS, NoA as a child class. Based on this definition, an Uncertainty-Aided Apportionment Network (UAAN) is proposed to classify complex pathological changes in IgAN, yielding two groups of predictions (corresponded to the parent and child classes) and their uncertainty. This uncertainty indicates the confidence coefficient of the prediction of the proposed network, which can be further applied to data reconstitution, including seeking missing annotations, mislabeled, and hard samples. With this indicator, the uncertain annotations can be picked up, rechecked, and analyzed by experts. To better illustrate the mechanism of UAAN and the

interpretability, heatmaps from different layers are visualized using Gradient weighted Class Activation Mapping (Grad-CAM) [10]. Besides, uniform manifold approximation and projection (UMAP) [11] is applied to demonstrate the data distribution before and after data reconstitution. Experimental results on the Warwick-QU dataset [12] and our in-house renal pathology dataset are reported, with a patch-wise and a slide-wise evaluation respectively.

The main contributions of our work are:

- We have introduced a scheme for fine-grained visual recognition in kidney pathology, aiming to detect complex pathological changes in IgAN.
- We have proposed a focal instance structural similarity loss to improve segmentation performance by assessing the structural integrity through the instance.
- We have designed an effective architecture for fine-grained classification and uncertainty assessment, with detailed ablation experiments and heatmap visualization.
- Comprehensive experiments have been conducted on multi-levels (patch-wise and slide-wise) to prove the effectiveness of our proposed network.

2. Related work

2.1. Loss functions for accurate segmentation

The common loss functions mainly consist of Dice coefficient loss, cross-entropy loss, Jaccard loss, and focal loss [13]. In addition to the well-known losses, Tversky loss [14] was proposed as an extension to the Jaccard loss, which restricted false positive and false negative rates through hyperparameters α and β. Lovasz loss combined Jaccard loss with Lovasz extension to find the minima of the submodular function [15]. Structural similarity was also introduced as an optimization function for segmentation based on sliding windows [16]. However, existing studies for segmentation losses barely assess the predictions through instance level [17]. Specifically, the

loss function should be designed considering each object within the image respectively, rather than adopting the same strategy for all the objects. In other words, the penalty of missing a small object should be higher than missing an equivalent area within the large object.

2.2. Fine-grained visual classification

The fine-grained visual classification (FGVC) was first defined to classify different species of birds from images with a single object [18,19]. It could be divided into region-based and feature expression-based methods. The region-based approaches [20,21] predominantly involved independent locations and fine-grained feature learning (forcing the classification model to learn discriminative features within specific regions). The feature expression-based methods normally required complex architectures or custom operations. Wu et al [22]. built a multi-path model consisting of M + 1 branches (M is the number of stain modalities), aiming to extract the complementary features from different stain modalities of a certain image. Ji et al [23]. incorporated convolutional operations along the edges of the tree structure and used the routing functions in each node to determine the root-to-leaf computational paths within the tree. However, due to the complex and various background samples, methods on natural images could barely achieve satisfactory performance in pathological studies.

3. Method

3.1. Overview

Given a whole slide image $\mathcal{H} \in R^{W \times H \times 3}$ that can be divided into $N_{\mathcal{P}}$ large patches $\mathcal{P} \in R^{\mathfrak{w} \times \mathfrak{h} \times 3}$ and binary mask $\mathcal{M} \in R^{\mathfrak{w} \times \mathfrak{h}}$, the target bounding boxes $\mathcal{B} \in R^{n_I \times 4}$ are first extracted to train the segmentation module $\mathcal{G}_\omega$. Optimized by the proposed Focal Instance Structural Similarity loss ($\mathcal{L}_{FISS}$), $\mathcal{G}_\omega$ aims to locate all types of the glomeruli for fine-grained classification. Then the classification module $\mathcal{D}_\theta$ is trained to predict the categories $\mathcal{J}$ of all segmented objects and produce the corresponding confidence coefficient $\mathcal{C}$ as

$$J, C = D(\mathcal{G}(\mathcal{P}, \mathcal{B}; \omega), \mathcal{F}, \mathfrak{l}_{CSN}; \theta) \tag{1}$$

where $\mathcal{F}$ denotes the fixed samples, $\mathfrak{l}_{CSN}$ indicates CSN index lists (details can be found in 3.3.1), ω and θ are the parameters of the segmentation module $\mathcal{G}_\omega$ and classification module $\mathcal{D}_\theta$. During the inference, the glomeruli are extracted by the segmentation model, instead of ground truth annotations.

3.2. Accurate segmentation of glomeruli using FISS

In this study, the segmentation module $\mathcal{G}_\omega$ is modified based on SegNet [24], introducing the group normalization and leaky relu. Currently, most segmentation losses are designed based on overlap measurements, while it makes the network only focus on the correct ratio of predicted pixels to the ground truth. To address this issue, a compound loss (DL+FISS) that considers overlap and regional structural similarity is introduced, which can be easily adopted in various architectures. The FISS loss combines focal loss L_F with instance structural similarity (ISS) loss L_{ISS}

$$L_{FISS} = \alpha \cdot L_F(p, g) + \beta \cdot L_{ISS}(p, g, \mathcal{B}) \tag{2}$$

where α and β are constraint weights, p and g represent the prediction and ground truth, $\mathcal{B}$ indicates the bounding box of each target in the ground truth. Since we want to balance the constraint from L_F and L_{ISS}, we set both α and β to 1, which was also proved appropriate in our initial pilot study. The Instance Structural Similarity (ISS) loss is inspired by Wang et al. [25], aiming to assess the structural integrity of each target

$$L_{ISS} = 1 - (ISS_P + ISS_N), \tag{3}$$

where ISS_P is the instance similarity of the image with Region of Interests (ROI) and ISS_N indicates the similarity of background samples (images without ROI). For images with multi-objects, ISS first evaluates the similarity between the prediction and the ground truth of each object. Then, a list that includes similarity indexes of instances is acquired, e.g., list $[I_0, I_1, I_2, I_3, I_4, I_5]$ will be acquired when there are six instances within the input image (shown in Supplement-Fig. 2). Let n_I be the number of instances in an input image x, μ_i, σ_i be the mean and variance of i-th instance within the x, the ISS_P can be given through

$$ISS_P = \frac{1}{n_I} \sum_{i=1}^{n_I} \frac{\left(2\mu_{p_i}\mu_{g_i} + c_{1_i}\right)\left(2\sigma_{p_i g_i} + c_{2_i}\right)}{\left(\mu_{p_i}^2 + \mu_{g_i}^2 + c_{1_i}\right)\left(\sigma_{p_i}^2 + \sigma_{g_i}^2 + c_{2_i}\right)}, \tag{4}$$

where p_i and g_i indicate the region of the ith instance extracted from the prediction and the ground truth of x. Instances that are smaller than the sliding window during similarity calculation are applied with zero paddings to ensure the kernel can slide across the object. Image without ROI is divided into λ patches (we set λ to 4) to compute ISS_N instead of evaluating the similarity directly

$$ISS_N = \min\left[SSIM(G_1, P_1), \ldots, SSIM(G_\lambda, P_\lambda)\right]. \tag{5}$$

By combining L_{ISS} and L_{Focal}, it will focus on the imbalanced samples while maintaining the structural integrity, which can output smoother boundaries and fewer false negative samples. Further exploration of FISS is illustrated in Supplement-3.

3.3. Fine-grained lesion classification via UAAN

This section introduces the details of UAAN ($\mathcal{D}_\theta$), including hierarchical structure design, feature apportionment mechanism, uncertainty aided data reconstitution, and solutions for imbalanced dataset.

3.3.1. Hierarchical structure design

Deep networks can achieve superior results when categories are independent and identically distributed. However, due to the complex pathological changes in IgAN, lesions cannot be well classified through normal solutions. In this section, we design a hierarchical architecture for fine-grained classification, with the parent branch B_p for classifying classes with large variation and child branch B_c for dividing class with small variation (shown in Fig. 3).

The backbone of UAAN is inspired by the Inception-ResNetV2 [26] by introducing Group normalization, PReLU [27] (to strengthen the generalization ability), and basic convolution block (including convolution layer, GN, and PReLU). After the backbone, feature maps are separately transported to two sub-branches (B_p and B_c) through CSN index list $\mathfrak{l}_{CSN}$, which is acquired during the data preparation with its elements representing the index of CSN samples in current batch. Assume a batch of N images containing m CSN samples, then $B_c \in R^{m \times w' \times h' \times c'}$, $B_p \in R^{N \times w' \times h' \times c'}$, where w', h', c' is the number of width, height, and channel in the last feature map of the backbone. This unique mechanism requires that each batch should include at least one CSN sample.

After the backbone of UAAN, the B_p and B_c are performed to give parent and child class predictions, respectively. In addition, a feature apportionment mechanism is proposed between these two branches to transfer valuable information of subcategories. In both B_p and B_c, the global maxpooling layers (GMP) are applied to extract the most intense response in each channel in the last feature map. And the dropout layers are implemented to force the network to learn more robust features that are useful in conjunction with different random subsets of other neurons. Details of the two branches can be found in Fig. 4.
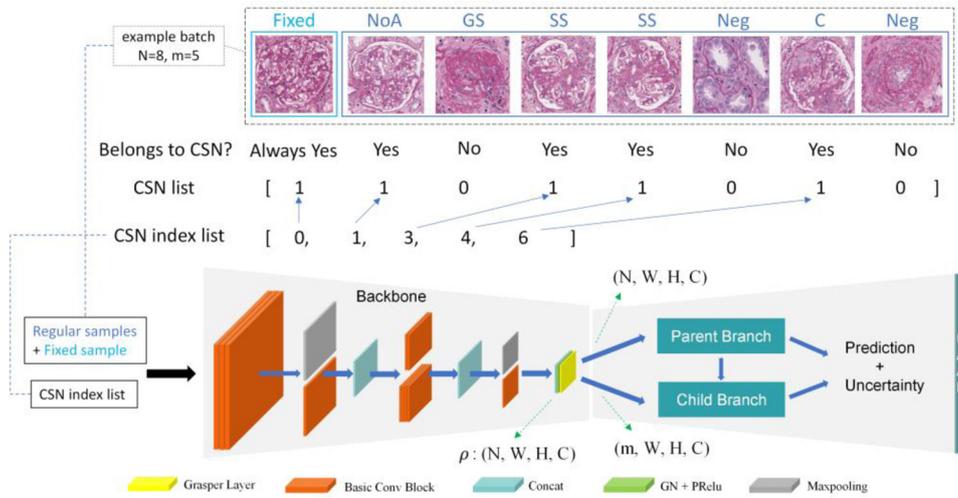
Fig. 3. The schematic diagram of the uncertainty-aided apportionment network. Each input mini-batch includes regular samples, a fixed sample, and a CSN index list.
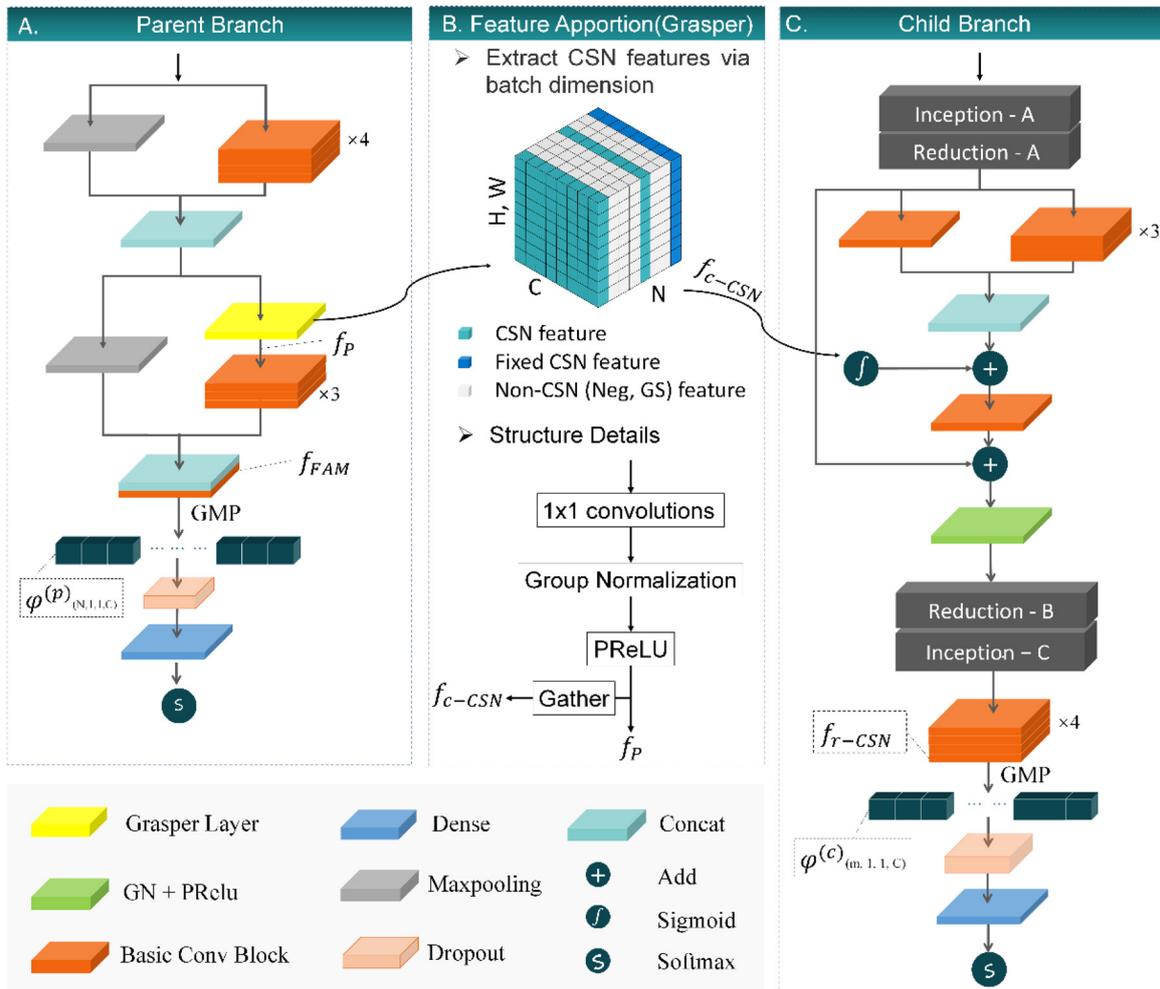

Fig. 4. The schematic diagram of the specific layers of the UAAN is shown in A. Parent branch. B. Feature apportion layer. C. Child branch.

3.3.2. Feature apportionment mechanism

Feature apportionment plays a crucial role in UAAN and is performed through the Grasper layer (Fig. 4B), aiming to extract CSN common features in the latent feature space. There are 5 classes: Neg, GS, C, SS, NoA, with a small inter-class variance between C, SS and NoA, and large inter-class variance between Neg, GS and CSN (the union of C, SS and NoA).

The gather operation in the Grasper layer extracts the certain feature maps of CSN, based on l_{CSN}. Since B_p is designed to classify Neg, GS, and CSN (here C, SS and NoA are regarded as a sin-

gle parent class), the layers in B_p aims to learn discriminative features among Neg, GS, and CSN samples. Therefore, the CSN feature in B_p can be regarded as the common feature of C, SS, and NoA (f_{c-CSN}) that rarely exists in GS and Neg samples (e.g., capillary loops, mesangial regions). On the contrary, since B_c aims to classify the C, SS, and NoA, layers in B_c are optimized to extract discriminative features of child classes. Therefore, the utilization of common and discriminative features can be promoted by transferring f_{c-CSN} from B_p to B_c through a sigmoid activation layer. This mechanism will help the network not only be trained with the discriminative features but also make use of common features.

3.3.3. Uncertainty aided data reconstitution

The hard samples in this study are samples that include many resembled features (or similar morphological characteristics) among C, SS, and NoA. For instance, the mesangial proliferation and crescent may co-exist in a single glomerulus, and experts usually give out their subjective judgment according to the severity degree. Due to these hard samples, even a large amount of data is provided, the network training can barely get good results. This is because including these samples in the training set makes it difficult for the network to extract features that are highly related to the real sub-category. In natural image analysis, these 'unreliable' and hard samples are usually tracked, relabeled, and augmented to enhance the capacity of computational modules. However, it is unaffordable for biomedical image analysis due to the heavy cost of expert annotation. We hold that some hard samples do not help to train a robust module, which leads to a negative swing of classification boundaries.

To bridge this gap, an uncertainty assessment is proposed to output the confidence coefficient of the prediction as to remove 'unreliable' and hard samples from the raw training set, calculated from the logits of the softmax layer $S^{(p)}(S^{(c)})$ and vectors of the GMP layer $\varphi^{(p)}(\varphi^{(c)})$ in both B_c and B_p. Deep networks can be regarded as powerful feature extractors that extract multi-level features in different layers. The dense layer and global pooling (GP) layer are in charge of dimension reduction, while the prediction is usually given in the last dense layer by weighing up the "evidence" from the previous layer. Because the input of the softmax layer is derived from the representation of a high dimensional feature vector, we consider the output of the softmax layer (representing a probability distribution over n different possible outcomes) as the estimated probability density P_e. Thus, the feature vectors $S^{(p)}(S^{(c)})$ are considered as the estimated probability density $P_e^{(p)}(P)_e^{(c)}$.

$$P_e^{(p)}, \ (P)_e^{(c)} = \phi(D(I_G, \ \mathcal{F}, \mathfrak{l}_{CSN}; \theta)), \tag{6}$$

where I_G refers to each glomerulus segmented by $\mathcal{G}_\omega$, ϕ indicates softmax activation, $P_e^{(p)}, \ (P)_e^{(c)}$ represent the estimated probability density of B_c and B_p. Besides, since the output $\varphi^{(p)}(\varphi^{(c)})$ of the last convolution layer can be regarded as the representation of I_G, the distribution of $\varphi^{(p)}(\varphi^{(c)})$ is close to the original distribution of the category that I_G belongs to. Thus, $\varphi^{(p)}$ and $\varphi^{(c)}$ are used to calculate the raw probability density function $P_r^{(p)}(P)_r^{(c)}$ to approximate the real probability distribution. Assuming the glomeruli feature distribution follows a Gaussian distribution, the feature vectors $\varphi^{(p)}$ and $\varphi^{(c)}$ are considered as a sampling from feature space. Therefore, the mean $\mu^{(p)}(\mu^{(c)})$ and variance $\sigma^{(p)}(\sigma^{(c)})$ of $\varphi^{(p)}(\varphi^{(c)})$, and the corresponding predictions can be calculated and substituted into

$$P_r(\mu, \sigma) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|k-\mu|^2}{2\sigma^2}}}{\sum_{k=0}^{K-1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|k-\mu|^2}{2\sigma^2}}} \tag{7}$$

to calculate the probability density function P, where K is the number of classes.

With the estimated probability density function P_e and the raw probability density function P_r, the uncertainty factor can be calculated using

$$U(P_r, \ P_e) = 1 - \frac{\sum_{k=0}^{K-1} \left(p_r^k - \bar{P}_r\right)\left(p_e^k - \bar{P}_e\right)}{\sqrt{\sum_{k=0}^{K-1} \left(p_r^k - \bar{P}_r\right)^2} \sqrt{\sum_{k=0}^{K-1} \left(p_e^k - \bar{P}_e\right)^2}}, \tag{8}$$

where P_r and P_e are denoted as

$$\begin{aligned}\bar{P}_r &= \frac{1}{K} \sum_{k=0}^{K-1} p_r^k \\ \bar{P}_e &= \frac{1}{K} \sum_{k=0}^{K-1} p_e^k\end{aligned}. \tag{9}$$

With this uncertainty factor, data reconstitution can be introduced to pick out and remove the 'unreliable' and hard samples. During the training procedure, the model was first trained from scratch on the raw dataset D_r, and then inferenced on D_r to conduct data reconstitution. The annotation is considered credible if there exists a strong correlation between P_r and P_e, and vice versa. Based on the uncertainty index, we picked up 2.33% NoA, 35.91% SS, 15.15% C, and 6.42% GS in total (when the uncertainty threshold was set as 0.5). Then, those "unreliable" samples are removed from D_r to reconstitute the training set D_c with the convinced samples. It is of note that the evaluation dataset remains unchanged during the entire period.

3.3.4. Solutions for the imbalanced dataset in classification

Details of imbalanced solution is presented in Supplement-4.

4. Experiments and results

4.1. Datasets and training strategies

The datasets in this study include Warwick-QU and our in-house Renal Pathological Whole Slide Images (RP-WSIs). The Warwick-QU dataset is used to assess the effectiveness of FISS loss (since segmentation of circular glomeruli is a relatively simple task compared to segmentation of various glands). The RP-WSIs is used to assess the effectiveness of. Details of these two sets are shown in Supplement-5.

Warwick-QU Dataset: The Warwick-QU [12] includes 85 training images and 80 test images, with the outer margin of glands annotated. The size of these images is 775 × 22 pixels.

Renal Pathological Whole Slide Images (RP-WSIs): 400 Periodic acid–Schiff stained WSIs of patients with biopsy-proven IgAN are selected from the National Clinical Research Center of Kidney Diseases in Jinling Hospital, ranging from 10,000 × 30,000 to 30,000 × 50,000 pixels. Essentially, 300 WSIs were used for training and validation and 100 WSIs were used for testing [28]. Details of preprocessing and training strategies can be found in Supplement-6.

4.2. Evaluation scheme and metrics

The evaluation scheme is divided into patch-wise and slide-wise. Patch-wise experiments aim to prove the capacity of FISS loss and UAAN, while those slide-wise are designed to demonstrate the comprehensive performance of the proposed scheme.

4.2.1. Patch-wise evaluation

(a) Evaluation of FISS Loss aims to illustrate the effectiveness of the FISS loss without considering any other interferences such as network architecture, tricks, data augmentation, and unique training strategies. We implemented ablation experiments without changing the network architecture to give. The participated loss functions include Dice coefficient loss (DL), Cross-Entropy loss (CE), Focal loss (FL), Tversky loss, Lovasz loss,

SSIM loss, Instance Structural Similarity loss (ISS), and Focal Instance Structural Similarity loss (FISS). Assume P as the prediction given by the segmentation network and G represents the ground truth, the segmentation task is assessed through Dice coefficient.

$$Dice = \frac{2|P \cap G|}{|P| + |G|}. \tag{10}$$

All these losses are trained on the same network without any architectural modification.

(a) Evaluation of UAAN consists of five parts. Firstly, we adopt ablation experiments to analyze the effectiveness of different techniques in UAAN, including feature apportionment and uncertainty-guided data reconstitution. All these models are trained from scratch and optimized by weighted cross-entropy loss with the same parameters. Secondly, we explore the performance of different networks to illustrate the capacity of the proposed UAAN on the raw dataset D_r (without data reconstitution) and convinced dataset D_c. Assume K as the number of classes, N as the number of whole test images. The micro-accuracy and macro-accuracy are reported as:

$$Acc_{micro} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Acc_{macro} = \frac{\sum_{i=0}^{K} \frac{TP_i}{N_i}}{K}. \tag{11}$$

where N is the total number of tested images and K is the number of classes, TP_i, N_i are the number of true positive, number of samples corresponding to the ith category, respectively. Thirdly, we visualize heatmaps of comparison models and that of the different layers in UAAN by Grad-CAM [10]. Heatmaps of the last convolution block in the parent branch f_{FAM}, common features of CSN f_{c-CSN}, and refined feature map of CSN f_{r-CSN} are visualized. Fourthly, we expurgate the "unreliable" data under different uncertainty thresholds from D_r and perform a second review by experts to explore the influence of different uncertainty thresholds. At last, the data distribution in latent space before and after uncertainty-aided data reconstitution is visualized by uniform manifold approximation and projection (UMAP) [11], calculated by vectors from GMP and Dense layer.

4.2.2. Slide-wise evaluation

The overall performance of the proposed scheme on whole slide images is assessed through the commonly used metric Average Precision (AP) and mean Average Precision (mAP). In this experiment, the ground truths (bounding boxes with classes tag) are generated through glomeruli boundaries given by experts. During the evaluation, the segmentation is first conducted to locate all glomeruli (stage-1), followed by the classification (stage-2). The inputs of classification are given by cropping the minimum bounding boxes of segmented glomeruli (given from stage-1). It is of note that all segmented objects except tiny objects (area less than 100 pixels) will be transferred to UAAN including incorrect-segmented samples. Then, the classification module produces category predictions of these cropped rectangles, and the probability given from stage-2 is regarded as the confidence threshold in mAP computations. Objects with a certain IoU threshold are considered as true positive samples (threshold 0.5 in our experiments). We compared the AP$_{50}$ (mean average precision with 0.5 IoU thresholds) of each subcategory among our scheme and various detection methods, including Mask R-CNN [8], FCOS [9], Mask Scoring R-CNN [29], and Cascade R-CNN [30]. All these models are trained from scratch without any pre-training.

Table 1
Ablation study of UAAN.

Feature Apportionment	Uncertainty	Acc_{micro}	Acc_{macro}
×	×	0.8960	0.8677
×	✓	0.9132	0.8935
✓	×	0 0.9294	0.8723
✓	✓	0.9517	0.9146

Table 2
Results of fine grained glomeruli lesion classification.

Model	Acc_{micro}-D_r	Acc_{micro}-D_c	Acc_{macro}-D_c
ResNet-50 [31]	0.8495	0.8717	0.7773
ResNeXt-50 [32]	0.8294	0.8474	0.6568
DenseNet-121 [33]	0.8635	0.8817	0.7757
EfficientNet-B6 [34]	0.9092	0.9235	0.7899
Inception-ResNet-v2 [26]	0.8507	0.8830	0.7396
Mean Teacher (70% annotation) [35]	/	0.7605	0.6067
DCL [36]	0.8176	0.8373	0.6770
SE-ResNeXt-101 [37]	0.8943	0.9106	0.7638
DenseNet+LSTM-GCNet [38]	0.9120	0.9295	0.8263
UAAN (ours)	0.9294	0.9517	0.9146

4.3. Results

4.3.1. FISS evaluation

Experimental results on Warwick-QU and RP-WSIs-P datasets for FISS loss are reported in supplement-7 through Dice coefficient score.

4.3.2. Uncertainty-aided apportionment network evaluation

Ablation Study of UAAN. The effectiveness of different techniques is presented in Table 1. It can be found that employing feature apportionment significantly improves micro-accuracy (with a 3.34% relative gain), while the uncertainty-aided data reconstitution effectively improves the macro-accuracy (with a 2.58% gain). When employing both feature apportionment and data reconstitution, there was a significant increase in both micro (5.57%) and macro (4.69%) accuracy compared with the baseline module.

Comparison of Fine-Grained Classification. UAAN is highly capable when dealing with fine-grained classification and outperforms all competing architectures at this task (Table 2).

The proposed UAAN achieves the best performance on both D_r and D_c with 95.07% micro-accuracy (3–11% relative gain) and 91.28% macro-accuracy (12–26%), respectively. It surpasses nearly 4% than EfficientNet-B6, 5.2% than SE-ResNeXt-101, 8% than Inception-ResNet-V2 and DenseNet-121. Meanwhile, it demonstrates that by reconstituting the training set through uncertainty factor, the performance of all comparison models has been significantly improved, with 1–4% micro-accuracy gain.

Visualization of heatmaps. Heatmaps of UAAN and competing networks are shown in Fig. 5 using Grad-CAM. The final prediction of UAAN is marked by black bounding boxes and the pathological changes in GS, C, and SS are annotated with yellow curves, while the "None" in f_{r-CSN} indicates the heat map of these two categories is not available in the child branch. It is obvious that heat maps of f_{c-CSN} (marked by the green rectangle) are highly related to the CSN common features, highlighting the capillary loops, intrinsic cells, and mesangial matrix. During the training procedure, these features are filtered by the grasper layer and conveyed to the child branch to enhance the common information.

The highlight regions in f_{FAM} straightly reflect where the network is "focusing", which directly affects the predictions among Neg, GS, and CSN. For instance, the highlight regions in $f_{FAM-NoA}$, f_{FAM-SS}, and f_{FAM-C} completely focus on the capillary loops, while the f_{FAM-GS} focuses on the mesangial region and $f_{FAM-Neg}$ concentrates on tubules and arteriole. Relevantly, these structures are ex-
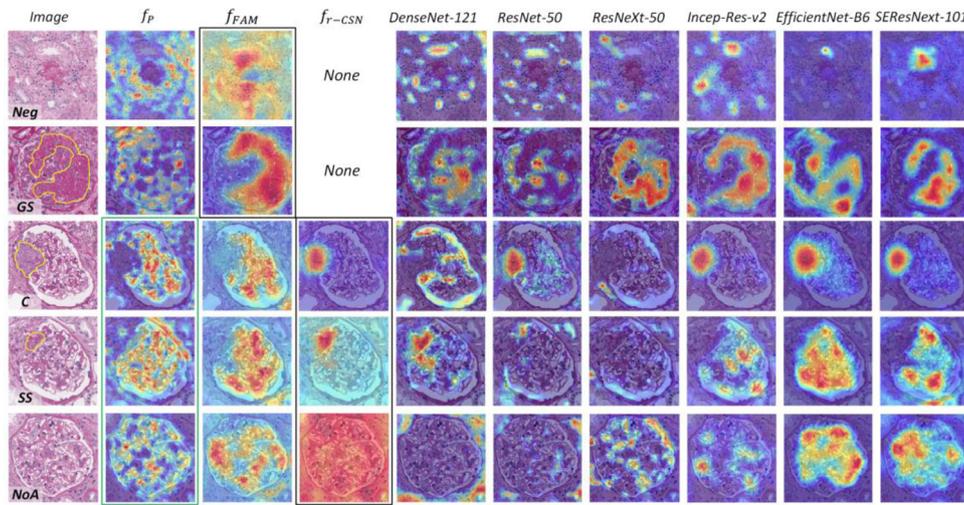
Fig. 5. Heatmap visualization of comparison models and different stages in UAAN. The heatmaps of the final prediction of UAAN are marked by black boxes and heatmaps of f_{c-CSN} are marked by the green rectangle.
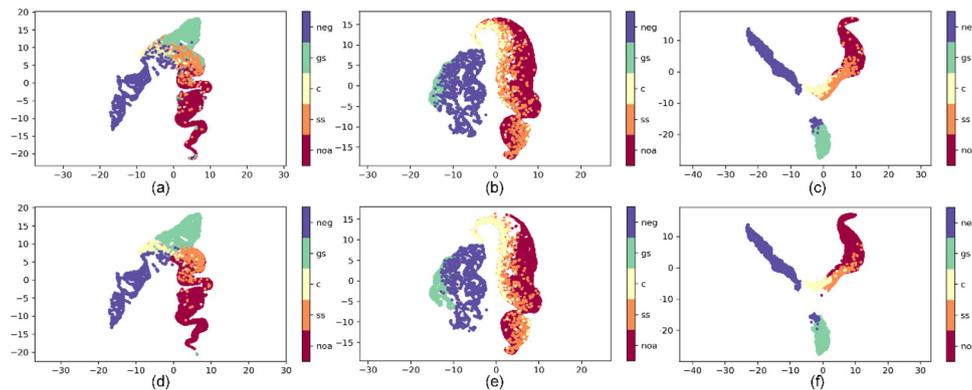

Fig. 6. Visualization of data distribution before (a–c) and after (d–f) data reconstitution via UMAP. (a) and (d) adopt the estimated probability vector from the last dense layer, (b) and (e) refer to raw probability vector from the GMP layer, (c) and (f) include both estimated and raw probability vector.

actly the main characteristics of distinguishing Neg, GS, and CSN by pathologists. Since the output of B_{child} only contains C, SS, and NoA, heat maps of Neg and GS are not available in f_{r-CSN} and are marked with *None* in Fig. 5. When observing f_{r-CSN}, the highlight areas appear high variance, with $f_{r-CSN-C}$ focusing on the proliferating cells in the Bowman's space, $f_{r-CSN-SS}$ concentrating on the increased mesangial matrix and collapsed capillary loops, and $f_{r-CSN-NoA}$ paying attention to global information (the network needs to check whether there exist pathological changes in each corner within NoA samples).

Ablation Study of Uncertainty Thresholds. The ablation study of different uncertainty thresholds is illustrated in Supplement-8.

Effectiveness of Data Reconstitution Fig. 6 (a–c) shows the data distribution before reconstitution, which is disordered, with certain NoA, SS, and C densely mixed. From observation in Fig. 6 (e–f), employing data reconstitution via uncertainty factor tends to tear off those mislabeled or hard samples in feature space, especially for SS, NoA, and C. While for those low variance categories, such as GS and Neg, are less affected. The distribution map in Fig. 6(f) indicates that the combination of probability vectors from GMP and dense layer exerts the clearest boundary, and the correlation between these vectors builds a bridge for uncertainty index.

4.3.3. Slide-wise evaluation

Table 3 presents the results of the slide-wise evaluation for fine-grained glomeruli detection on 100 whole slide images. The

Table 3
Performance of glomeruli detection.

Model	AP_{NoA}	AP_{SS}	AP_{GS}	AP_C	mAP
Mask R-CNN [8]	0.705	0.103	0.707	0.240	0.439
FCOS [9]	0.789	0.280	0.751	0.336	0.539
MS R-CNN [29]	0.674	0.110	0.555	0.230	0.392
Cascade R-CNN [30]	0.785	0.165	0.780	0.194	0.481
Our Method (full 2-stage)	0.780	0.394	0.718	0.575	0.613

visualization results of slide-wise evaluation are presented in Supplement-9.

Experimental results show that powerful detectors such as Mask R-CNN, FCOS, MS R-CNN, and Cascade R-CNN present weaker performance than the proposed method (0.613 mAP), with an 8–22% reduction of mAP. To further analyze the evaluation results, the Wilcoxon-Mann-Whitney test is adopted. Statistical results show that there is no significant differences (p-value> 0.05) between the proposed method and the above two when identifying GS (p-value $= 0.0727$ for FCOS & ours, p-value $= 0.0619$ for Cascade R-CNN & ours) and NoA (p-value $= 0.0895$ for FCOS & ours, p-value $= 0.0978$ for Cascade R-CNN & ours) samples. Though FCOS and Cascade R-CNN appear similar performance to the proposed method in identifying NoA and GS samples, they show poor capacity in distinguishing SS and C. By contrast, the proposed method shows balanced performance and can effectively detect all kinds of glomerular lesions, especially for SS and C.

5. Discussion

This section provides discussions on the proposed method, including (1) quantitative analysis of segmentation losses (2) fine-grained classification in pathological images (3) uncertainty assessment in the deep convolutional neural network, and (4) the limitation when current detection-based methods dealing with FGVR tasks.

Segmentation Loss Functions. We have compared different loss functions with the FISS loss by evaluating the Dice score on simulated experiments, Warwick-QU and renal pathology datasets, respectively. Supplement-Fig. 3 shows that the proposed compound loss is the most sensitive one to different sizes of targets. Supplement-Table 3 indicates the model trained with the proposed compound loss can achieve better performance among mainstream loss functions, including Dice loss, cross-entropy loss, focal loss, Tversky loss, Lovasz loss, and SSIM loss. Although training with Dice loss and other aided losses could effectively suppress false-positive samples, it also leads to more false-negative samples, while applying DL+FISS loss can achieve an equilibrium, performing better boundaries and smooth predictions. Although there is a slight reduction of the structural integrity, employing DL+FISS actually achieves better performance compared to using the DL+ISS only.

Fine-Grained Classification in Pathological Images. The ablation study in Table 1 reminds researchers to make better use of common features between different categories, instead of only forcing the network to focus on discriminative regions. As Table 1 shows, feature apportionment mainly helps the network to distinguish NoA from C and SS, while classifying C and SS still requires a cleaner data constitution. Through the reconstitution of the training set, the classification module can find a better boundary to distinguish fine-grained classes. By integrating uncertainty-aided data reconstitution with a feature apportionment mechanism, both the micro and macro accuracy are significantly improved. In Table 2, the proposed method has gained a competitive edge over other basic models, with 95.07% micro-accuracy (3–11% relative gain) and 91.28% macro-accuracy (12–26%). This huge gap has indicated the poor classification ability of current competitive models [26,31–34,36,37] when dealing with data imbalanced issue. Compared with most fine-grained classification methods, the proposed UAAN has several advantages: (1) It has not required a large number of branches or generative models (2) It has considered both discriminative features and common features between subcategories, instead of focusing on the discriminative part only (3) It has a streamlined structure and can be further improved by adding complex operations such as atrous spatial pyramid pooling and attention module. (4) It has not required any part annotations or subnetworks.

Interpretability of UAAN. The visualization of heat maps at different layers has proved the reasonability of our architectural design. Besides, we have presented heatmaps of the comparison models in terms of different glomerular lesions. Among these five categories, SS has been found as the most difficult class for recognition, with only UAAN and DenseNet-121 focusing on the right place. As Fig. 5 illustrated, even some state-of-the-art models exert high accuracies (such as SE-ResNeXt-101, EfficientNet-B6), their attention regions have not been fully concentrated on the pathological changes. In contrast, UAAN's attention is much closer to pathologists, performing high interpretability of the prediction.

Uncertainty in Deep Convolutional Neural Network. The uncertainty is quite important in deep convolutional neural networks (especially in medical applications) since it reflects the confidence coefficient of the current predictions. In our study, we have introduced the uncertainty assessment by calculating the correlation between two probability densities. Although our assessment has not provided a strict uncertainty calculation using Bayesian inference, it can still estimate the confidence coefficient of the predictions to some extent. Besides, it has also been used for selecting low-annotation data and hard samples, which has not required multiple forward propagations. The ablation experiments in Supplement-Table 5 show that, with the increase of uncertainty threshold, fewer samples have been selected while the ratio of the mislabeled samples is higher. With detailed inspection by three senior pathologists, 36% and 55% among the mislabeled images have been corrected as C and SS (in case of $U_{thresh} = 0.5$). This has proved our assumption that the output of the GMP layer is more correlated to the real distribution and can be used to simulate the raw probability density function. If there is no obvious correlation between P_r and P_e, we consider the features are misrepresented in the last dense layer. That is, the output of the softmax layer represents an erroneous probability distribution due to the inaccurate feature representation process. It might be the reason why CNN+SVM is better than pure CNN in some previous studies [39]. Aided by this uncertainty assessment, significant improvements (1–4%) of micro-accuracy have been achieved by current state-of-the-art models [26,31–34,36,37] using data reconstitution.

Comparisons with junior pathologists. We compared the performance of UAAN with a junior pathologist on a pilot dataset (200 images per class). The labels of this pilot dataset were given by two senior pathologists, with the inter-observer agreement of Cohen's Kappa = 0.791, $p < 0.0001$. The junior pathologist achieved 89.4% accuracy compared to the 90.7% of UAAN's, while the UAAN (7.57 ms per glomerulus on Nvidia Tesla V100 16GB, batch size 32) is nearly 260 times more efficient compared to a junior pathologist (2 s per glomerulus), which shows the feasibility of conducting pathological diagnosis by neural networks.

Why two stages? Though the current state-of-the-art detection modules can achieve competitive performance for locating all types of glomeruli compared with the two-stage scheme, these modules cannot well classify the fine-grained lesions. One main reason is the resolution. Due to the scale of viewings, the detection module can only be well performed under 4X resolution (2.468 μm/pixel) with an input size of 2048 × 2048, however, under certain resolution the model cannot extract fine-grained features to distinguish fine-grained subcategories. Moreover, adding complex classification heads accounts for a huge module that leads to the shortage of GPU memory.

Limitations. Although our work has achieved superior performance among the current methods, it has some limitations. The proposed two-stage scheme suffers from a complex feature extraction procedure, and the performance of stage-2 (fine-grained classification) is highly related to that of the segmentation stage. For instance, the classification module cannot work well when two contiguous glomeruli cannot be segmented individually. In addition, the various annotation qualities, expert subjectivity, and dataset may affect the stability of data reconstitution. Indeed, external validation can be helpful for testing the generalizability of our proposed deep learning model, but at the moment we have no access to additional data resources. In lieu of constructing an external validation dataset, we will publish our implementation on opensource platform for reproducibility studies and external validation by other researcher groups.

6. Conclusion

In this study, we have introduced a comprehensive scheme for glomeruli lesion recognition, employing focal instance structural similarity loss and the uncertainty-guided apportionment network. This work is the first attempt to tackle the fine-grained visual recognition task in pathological image analysis and has achieved superior performance in IgAN whole slide images. Both the focal

instance structural similarity loss and uncertainty-aided apportionment networks are effective, resulting in more than 8–22% improvement of the mAP compared with current schemes. The proposed method provides a high-precision computational scheme for fine-grained lesion identification of IgA nephropathy in whole slide images, which helps pathologists make more objective and effective clinical diagnoses. For future work, it is of note that the detected lesions are not specific for IgA nephropathy, and one of our future research directions will be the transfer learning for the lesion's detection to other nephropathies with a similar presentation, e.g., lupus nephritis and diabetic nephropathy.

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.

Acknowledgments

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2022.108648.

References

[1] I.S.D. Roberts, H.T. Cook, S. Troyanov, C.E. Alpers, A. Amore, J. Barratt, F. Berthoux, S.M. Bonsib, J.A. Bruijn, D.C. Cattran, The oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility, Kidney Int. 76 (2009) 546–556.

[2] J.D. Bukowy, A. Dayton, D. Cloutier, A.D. Manis, A. Staruschenko, J.H. Lombard, L.C.S. Woods, D.A. Beard, A.W. Cowley, Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections, J. Am. Soc. Nephrol. 29 (2018) 2081–2088.

[3] B. Ginley, B. Lutnick, K.Y. Jen, A.B. Fogo, S. Jain, A. Rosenberg, V. Walavalkar, G. Wilding, J.E. Tomaszewski, R. Yacoub, Computational segmentation and classification of diabetic glomerulosclerosis, J. Am. Soc. Nephrol. 30 (10) (2019) 1953–1967 ASN. 2018121259.

[4] S. Sheehan, S. Mawe, R.E. Cianciolo, R. Korstanje, J.M. Mahoney, Detection and classification of novel renal histologic phenotypes using deep neural networks, Am. J. Pathol. 189 (2019) 1786–1796.

[5] G.S. Xie, X.Y. Zhang, W. Yang, M. Xu, S. Yan, C.L. Liu, LG-CNN: from local parts to global discrimination for fine-grained recognition, Pattern Recognit. 71 (2017) 118–131.

[6] L. Qi, X. Lu, X. Li, Exploiting spatial relation for fine-grained image classification, Pattern Recognit. 91 (2019) 47–55.

[7] X.S. Wei, C.W. Xie, J. Wu, C. Shen, Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization, Pattern Recognit. 76 (2018) 704–714.

[8] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[9] Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in: Proceedings of the International Conference on Computer Vision, 2019, pp. 9627–9636.

[10] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad–cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[11] L. McInnes, J. Healy, J. Melville, Umap: uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426, (2018).

[12] K. Sirinukunwattana, D.R. Snead, N.M. Rajpoot, A stochastic polygons model for glandular structures in colon histology images, IEEE Trans. Med. Imaging 34 (2015) 2366–2378.

[13] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[14] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: Proceedings of the International Workshop on Machine Learning in Medical Imaging, 2017.

[15] M. Berman, A. Rannen Triki, M.B. Blaschko, The Lovász-Softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4413–4421.

[16] S. Zhao, B. Wu, W. Chu, Y. Hu, D. Cai, Correlation maximized structural similarity loss for semantic segmentation, arXiv preprint arXiv:1910.08711, (2019).

[17] S.A. Taghanaki, K. Abhishek, J.P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: a review, Artif. Intell. Rev. (2020) 1–42.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, (2011).

[19] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 554–561.

[20] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based R-CNNs for fine-grained category detection, in: Proceedings of the European Conference on Computer Vision, 2014.

[21] S. Huang, Z. Xu, D. Tao, et al. Part-stacked CNN for fine-grained visual categorization[C], Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 1173–1182.

[22] B. Wu, X. Zhang, S. Zhao, L. Xie, C. Zeng, Z. Liu, G. Sun, G2C: a generator–to-classifier framework integrating multi-stained visual cues for pathological glomerulus classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 1214–1221.

[23] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, F. Huang, Attention convolutional binary neural tree for fine-grained visual categorization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10468–10477.

[24] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495.

[25] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[27] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[28] W.M. Association, World medical association declaration of helsinkie thical principles for medical research involving human subjects, Chin. J. Integr. Med. 2 (2001) 92–95.

[29] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask scoring R-CNN, in: Proceedings of the Computer Vision and Pattern Recognition, 2019, pp. 6409–6418.

[30] Z. Cai, N. Vasconcelos, Delving into high quality object detection, in: Proceedings of the Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.

[31] K. He, X. Zhang, S. Ren, S. Jian, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[32] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[33] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[34] M. Tan, Q.V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 6105–6114.

[35] A. Tarvainen, H. Valpola, Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. 30 (2017) 30.

[36] Y. Chen, Y. Bai, W. Zhang, T. Mei, Destruction and construction learning for fine-grained image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5157–5166.

[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[38] C. Zeng, Y. Nan, F. Xu, Q. Lei, F. Li, T. Chen, S. Liang, X. Hou, B. Lv, D. Liang, Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning, J. Pathol. 252 (2020) 53–64.

[39] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive review, Neural Comput. 29 (2017) 2352–2449.

Yang Nan obtained his MEng from Hunan University. He is currently a PhD student and research assistant in NHLI, Imperial College London. Before that, he was a senior researcher in PingAn Healthcare Technology Co. Ltd.

Fengyi Li received master degree from University of Zhejiang University in 2018. She was a research scientist with Pingan tech. Her research topic relates to computer vision, natural language process and causality learning. She is currently a researcher in Antgroup, China, engaging in intelligent decision and attribution analysis.

Peng Tang currently is a Phd student of Department of Informatic and CIIP research group at Technical University of Munich. He obtained the Bachelor degree from Shaoyang University in 2012 and Master degree from Hunan University in 2020. His-research interests include deep learning and medical image analysis.

Guyue Zhang received her Ph.D. degree in the School of Computer Science at Fudan University in 2018. Now she is a senior researcher in Ping An Technology. Her research topic relates to deep learning and computer vision, especially human detection and tracking, action detection, and medical image processing.

Caihong Zeng is a renal pathologist and professor of nephrology, Ph.D. supervisor of national clinical research center of kidney diseases, Jinling Hospital, Nanjing University School of Medicine. Her research focuses on kidney disease and application of artificial intelligence in renal pathology. She was a Training Committee member of Renal Pathology Society Committees.

Guotong Xie received his Ph.D. degree in the School of Computer Science at Peking University. He was a senior technical staff in IBM China Research Lab. His-research topic relates to real world evidence, cognitive decision support and medical imaging. He is currently a Chief Healthcare Scientist in Ping An Group.

Zhi-Hong Liu is professor of medicine, Academician of Chinese Academy of Engineering. Director, National Clinical Research Center of Kidney Disease, Jinling Hospital, Nanjing University School of Medicine. Her primary interest is in the field of kidney disease, with special interest in glomerulonephritis, diabetic nephropathy and rare diseases.

Guang Yang obtained his MSc and PhD from University College London. He is currently a tenured senior research fellow (UKRI Future Leaders Fellow) at NHLI, Imperial College London and an honorary senior lecturer at King's College London. He is the head of the Smart Imaging lab funded by UKRI, BHF and ERC.