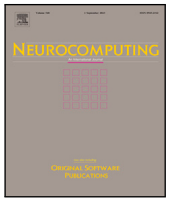
Neurocomputing 563 (2024) 126947

Contents lists available at ScienceDirect

Neurocomputing

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



Enhancing glomeruli segmentation through cross-species pre-training

Paolo Andreini [a,1], Simone Bonechi [a,b,1], Giovanna Maria Dimitri [a,*,1]

[a] Department of Information Engineering and Mathematical Science, Via Roma 56, Siena, 53100, Italy
[b] Department of Social Political and Cognitive Science, Via Roma 56, Siena, 53100, Italy

ARTICLE INFO

Keywords:
Deep Learning
Histopathology
Kidney
Semantic segmentation

ABSTRACT

The importance of kidney biopsy, a medical procedure in which a small tissue sample is extracted from the kidney for examination, is increasing due to the rising incidence of kidney disorders. This procedure helps diagnosing several kidney diseases which are cause of kidney function changes, as well as guiding treatment decisions, and evaluating the suitability of potential donor kidneys for transplantation. In this work, a deep learning system for the automatic segmentation of glomeruli in biopsy kidney images is presented. A novel cross-species transfer learning approach, in which a semantic segmentation network is trained on mouse kidney tissue images and then fine-tuned on human data, is proposed to boost the segmentation performance. The experiments conducted using two deep semantic segmentation networks, MobileNet and SegNeXt, demonstrated the effectiveness of the cross-species pre-training approach leading to an increased generalization ability of both models.

1. Introduction

Kidney biopsy is an essential diagnostic tool for several kidney diseases and its importance is increasing due to the rising incidence of renal disorders [1].

This is a procedure frequently performed by doctors, which involves extracting a small sample of tissue from a potential donor kidney for microscopic investigation [2]. The biopsy is an essential procedure, as it can provide valuable information about correct kidney functioning and help diagnose various kidney disorders [3–5] (e.g. Nephritis,[2] Nephrotic Syndrome,[3] kidney tumors, and other specific cancer pathologies related to kidneys). In addition, the histopathological analysis conducted on the tissue samples obtained through biopsy can also help in determining the cause of unexplained kidney function changes and guides treatment decisions, such as the choice of medications or the need for dialysis. Moreover, more related to our work, analyzing kidney biopsies is a crucial step in evaluating kidneys for transplantation [6]. Indeed, the biopsy allows to analyze glomeruli, which serve as functional units of the kidney's filtration process. The glomeruli, situated within each nephron of the kidney, are intricate clusters of tiny blood vessels. These specialized structures serve as the key filtering units responsible for removing waste materials and excessive fluids from the bloodstream. The presence of damaged or malfunctioning glomeruli can lead to kidney disease [7] and need to be considered in the evaluation of potential transplant candidates. In this context, the analysis of glomerular structure can provide important information about the health and working capabilities of the kidney [6]. Indeed, its structure is particularly critical for distinguishing between sclerotized and non-sclerotized glomeruli. The percentage of these types of glomeruli observed in a kidney tissue sample is used to calculate the Karpinski index [8], which is a fundamental indicator for assessing renal function and has significant value in evaluating potential kidney donors. In fact, analyzing glomeruli from the biopsy can provide insights concerning the cause of kidney disease and guide the selection of the best treatment options for the transplant recipient. The results of glomeruli evaluation can help ensure the success of transplantation by reducing the risk of post-transplant complications and improving the long-term outcome for the recipient patients [9,10]. As an example, in presence of conditions like focal segmental glomerulosclerosis, a disease related to drugs abuse, a biopsy is crucial to evaluate the proper kidneys functionality [11–13].

Usually, glomeruli identification and counting, a repetitive and time-consuming procedure, is carried out visually by an experienced clinician. Timely completion of this task is essential to assess the quality of the donor's kidney and increase the probability of a successful transplantation [14]. These observations motivate the interest in developing

* Corresponding author.
E-mail address: giovanna.dimitri@unisi.it (G.M. Dimitri).

[1] Equal contribution.

[2] Kidney's inflammation which could be chronic or temporary.

[3] Nephrotic syndrome is a broad term used to describe a collection of clinical symptoms that signify impaired kidney function. These symptoms may include excessive protein in the urine (proteinuria) and swelling in various body parts, known as edema.

https://doi.org/10.1016/j.neucom.2023.126947
Received 16 February 2023; Received in revised form 16 September 2023; Accepted 16 October 2023
Available online 18 October 2023
0925-2312/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

effective automated tools to support clinicians in this challenging task. For this reason, in this work, we propose a deep learning system for the automatic segmentation of glomeruli in kidney biopsy images.

Deep Learning (DL) has achieved tremendous success across various fields, including computer vision [15–17], biomedicine [18,19], and natural language processing [20,21]. In the medical field Convolutional Neural Networks (CNNs) became very popular and were used to develop many medical applications, from the segmentation of retinal fundus images [22] and skin lesions [23] to the analysis of radiological, magnetic resonance or computerized axial tomography images [24–26]. In particular, the use of these models offers a valuable contribution to clinical practice, reducing costs, and increasing repeatability, and accuracy. Nonetheless, the success of deep neural networks is usually strictly related to the availability of large sets of supervised training data. This limitation is particularly significant for the development of semantic segmentation networks where the need for pixel-level supervision, makes the collection of annotated data particularly time-consuming and costly.

The primary purpose of this study is to demonstrate the potential of cross-species transfer learning in overcoming the scarcity of pixel-level labeled images for glomeruli segmentation in histopathological images of human kidneys. Indeed, in this context, the proposed approach can be significantly valuable since most of the existing deep learning glomeruli segmentation models that can be found in literature, rely on in-house proprietary datasets that are not publicly available. In this paper, extending our seminal work [27], we have used two public datasets: the mouse kidney glomeruli datasets [28] and the HuBMAP dataset released for the Kaggle "Hacking the Kidney" challenge.[4] In particular, we propose a new segmentation approach based on cross-species transfer learning: a semantic segmentation network is first trained on a dataset of mice kidney tissue images and then fine-tuned on human data. The rationale behind this approach is that despite the difference between the two species (humans and mice), the morphological aspects of tissues share some similarities [29].

Indeed, human and mice glomeruli exhibit striking similarities in their anatomical features and functions [30]. Their structure is comparable in both species, characterized by a tuft of capillaries lined with endothelial cells and surrounded by podocytes. Moreover, they share a common purpose of blood filtration, playing a crucial role in upholding fluid and electrolyte equilibrium and regulating blood pressure. Several studies, carried out during the past years, highlight similarities between mice and human kidney tissue. For instance, [31] proposed a study on diabetic kidneys, evaluating both mice and human samples highlighting the presence of significant similarities between the two cases. Similarly, other studies related to diabetic pathologies have identified commonalities between the two species [32,33]. Interestingly, common features between mice and human tissues have been also found analyzing kidney tumors when local lesions and inflamed tissues in tertiary lymphoid-affected kidneys are present [34–36]. Motivated by these results, we hypothesize that a neural network could leverage the acquired knowledge on the mice images when applied on the human dataset, thereby enhancing its overall generalization capabilities. Indeed, our experiments confirm that by employing the proposed cross-species approach it is possible to obtain a significant performance boost, confirming the initial intuition. The experiments were carried out with two deep semantic segmentation networks, the MobileNet [37] and the SegNeXt [38]. The SegNext model is a state-of-the-art segmentation network that achieved top performances on a variety of common benchmarks [38], which motivated its choice in this study. Other models like the UNet [39] and the DeepLab [40], which was used in our seminal work [27], were discharged since they provided significantly lower performance compared to the SegNext in our preliminary experiments conducted on mice glomeruli segmentation. Additionally, with the purpose of exploring the feasibility of implementing a glomeruli segmentation model that can efficiently run on low-end hardware setups, we chose to utilize the MobileNet as an additional model. In fact, the SegNeXt and MobileNet may represent an interesting case study, since they have a significantly different number of trainable parameters that may influence their generalization capabilities and their effectiveness in the proposed cross-species pre-training approach. Indeed, MobileNet V3 has only 3.3M parameters, while SegNeXt, in the implementation used in this work, comprises 13.9M parameters. The structure of the paper is as follows: Section 2 presents a review of relevant literature. Section 3 provides an overview of the datasets used in this study, while Section 4 introduces the two segmentation network architectures that were employed. The experimental setup is described in Section 5, and finally, the conclusions and future developments are discussed in Section 6.

2. Related works

Recently, Deep Learning (DL) has sparked a genuine revolution in computer science. Numerous fields have benefited from the use of DL algorithms, including bioinformatics, natural language processing, and object detection. Among the many successful fields of applications of DL one of the most prominent is semantic segmentation where in just a few years the performance impressively improved [41–43]. Image semantic segmentation aims at making dense predictions classifying each pixel in an image into a predefined set of categories. DL algorithms can be used to perform this task producing as output a segmented image where each pixel is assigned to a class label. Thus this can provide a deeper understanding of the structure and content of the image itself. Several works have employed deep semantic segmentation models in histopathology (for a more comprehensive overview, refer to [44]). In the subsequent sections, a review of various studies on glomeruli segmentation will be provided. Specifically, Sections 2.1 and 2.2 revise some works focused on the segmentation of human and mice glomeruli, respectively.

2.1. Human kidney Glomeruli segmentation

Several DL applications have been recently proposed to automatically evaluate kidney tissues for transplant. For instance, a deep learning model to identify and segment glomerular structures in human kidney biopsies has been presented in [13]. In this study, a set of 275 images from patients with renal diseases were analyzed using a multi-class CNN to segment sclerotized and non-sclerotized glomeruli. Moreover, in [12] the authors presented a segmentation approach for human glomeruli in a new dataset; the use of different staining procedures (Masson and CD10) was explored to enhance the glomerular structures in the images, showing promising results. In [45], the authors applied two common CNN networks, the UNet and SegNet, for the segmentation of glomeruli in human kidney histopathological images. Similarly, in [46] an instance segmentation network, the Mask R-CNN [47], was employed for glomeruli segmentation and classification. The experiments were performed on a dataset, composed of 26 kidney biopsies collected from 19 donors, provided by the Department of Emergency and organ Transplantations of the University of Bari. Furthermore, in [48] the authors compared different Deep Learning approaches for the semantic segmentation of glomerular structures obtaining good performance employing a UNet based architecture. Finally, more related to this work, several previous studies have used the HuBMap dataset for various purposes. For example, in [49] the authors used the dataset to classify sclerotic and non-sclerotic glomeruli. However, the dataset is more commonly used for glomeruli segmentation, and a number of different approaches have been proposed for this purpose in recent years. Unfortunately, a fair comparison between these approaches is not possible due to the difference in the experimental

[4] https://www.kaggle.com/competitions/hubmap-kidney-segmentation/data.

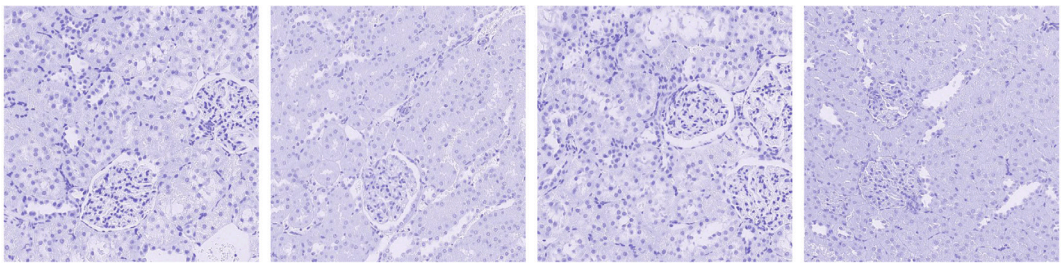


Fig. 1. Some examples of image tiles taken from the training set of the mice glomeruli dataset.

setup and in the evaluation protocols. For example, in [50] the authors used a UNet-based architecture for glomeruli segmentation on the HuBMap dataset. However, instead of using the official test set, they evaluated the model on a subset of the training set without specifying what are the images employed for the evaluation. Similarly, in [51], the HuBMap dataset was used to train a segmentation network but the official test set was not employed for testing. In [52] the authors achieved excellent performance using an ensemble of different models. However, the results are reported on the HuBMap private test set, making the comparison with other models infeasible if they are evaluated on the public test set.

2.2. Mice kidney Glomeruli segmentation

Only few works focus on the semantic segmentation of glomeruli in mice kidney tissue images. For instance, in [53] the mice glomeruli segmentation is performed using a standard image processing approach based on Gabor filters. Instead, a DL architecture is exploited in [54] where a semantic segmentation network with an AlexNet backbone is employed. Similarly, in [28] a CNN is used to segment mice kidney images, the model performs both detection and segmentation and it is evaluated separately for different staining procedures. The authors released the trained model and the dataset along with the Orbit Image analysis tool. Finally, more related to this work, in [27] the dataset released by [28] was employed to compare different deep architectures. The work presented in this paper extends [27], demonstrating that it is possible to exploit the pre-training on mice data to increase the performance of human glomeruli segmentation.

3. Datasets

In the following section, we will briefly describe the characteristics of the two datasets used in our experiments.

3.1. Mice Glomeruli dataset

The mouse glomeruli dataset used in this study was released together with the Orbit Image Analysis tool [28], an open-source software developed for the analysis of histopathological images. The dataset is publicly available[5] and comprises 88 Whole Slide Images (WSIs) with an average resolution of $10{,}000 \times 8{,}000$ pixels. More specifically, WSIs are images represented in a pyramidal structure composed of several sub-images at different resolutions. The histopathological images come from two different species (mice and rats) and were processed using various staining techniques including Hematoxylin (H), Eosin (E), Diaminobenzidine (DAB), Immune Chromogenic Reagent (FastRed), Periodic Acid-Schiff (PAS), and three variations of H and E. The dataset contains the supervision, consisting of manually annotated masks where the position of each glomerulus is defined at pixel-level, for about 21 000 glomeruli. The annotations are released in the SQLite database format and can be opened and exported using the Orbit software. In this study, we adopted the official dataset split, where the test set consisted of 32 images, the validation set comprised of 8 images, and the remaining images were used for training. Some image tiles selected from the training set are shown in Fig. 1.

3.2. Human Glomeruli dataset

The human kidney dataset used in this paper was released for the HubMap Kidney Challenge on Kaggle whose goal was to develop new tools for the identification of glomeruli in PAS-stained images. The dataset, which can be freely downloaded from the Kaggle website,[6] is composed of 11 fresh frozen and 9 formalin fixed paraffin embedded PAS kidney images. Each image is in tiff format with an average resolution of $36{,}000 \times 29{,}000$ pixels. Whereas the pixel-level annotations are released in a JSON file. In particular, each glomerulus is identified by the vertices of a polygon that encompasses it. The test set is composed of 5 images whose annotations are not publicly available: a quantitative evaluation of the results on the test images can be obtained by submitting the predictions on the challenge server. Because an official validation set has not been released, in this work three random images have been removed from the training set and used as validation. Some image tiles selected from the training set are shown in Fig. 2.

Additionally, each image of this dataset is associated with some anamnestic data of the patient. Statistics about the data distribution of the dataset are presented in Fig. 3. Specifically, we report the gender, the Body Mass Index (BMI), the age of the study participants, and the laterality (right or left kidney) associated with the samples. As we can observe from the figure, the dataset exhibits a significant balance in terms of sex and laterality, indeed it contains an equal distribution of males and females, as well as an equal representation of left and right kidneys. Moreover, we could also observe that most of the patients are between 50 and 60 years old.

4. Segmentation networks

In the following section, the deep segmentation network models employed in this work are briefly described.

4.1. MobileNet

MobileNets are a family of efficient neural network models specifically designed to have a reduced hardware footprint, so that they can be easily integrated into mobile and embedded devices. A common feature of these architectures is the use of depth-wise separable convolutions that allows to build lightweight deep neural networks. In particular, the MobileNet [55] architecture is designed to offer a trade-off between latency and accuracy. While the original MobileNet model was proposed for image classification, the architecture has been adopted also to perform other tasks like semantic segmentation. In particular,

[5] https://datadryad.org/stash/dataset/doi:10.5061/dryad.fqz612jpc.

[6] https://www.kaggle.com/competitions/hubmap-kidney-segmentation/data.

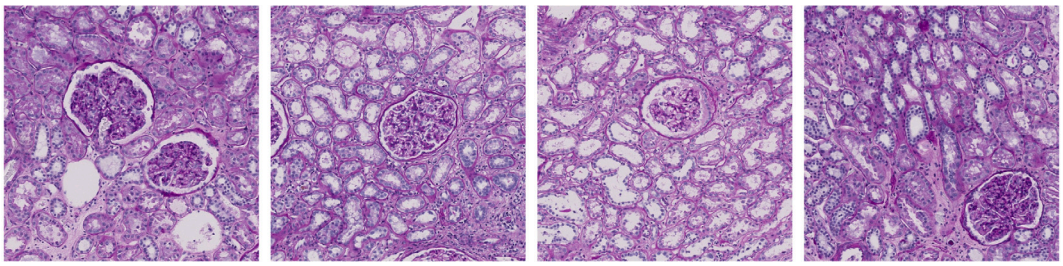


Fig. 2. Some examples of image tiles taken from the training set of the HubMap kidney glomeruli dataset.

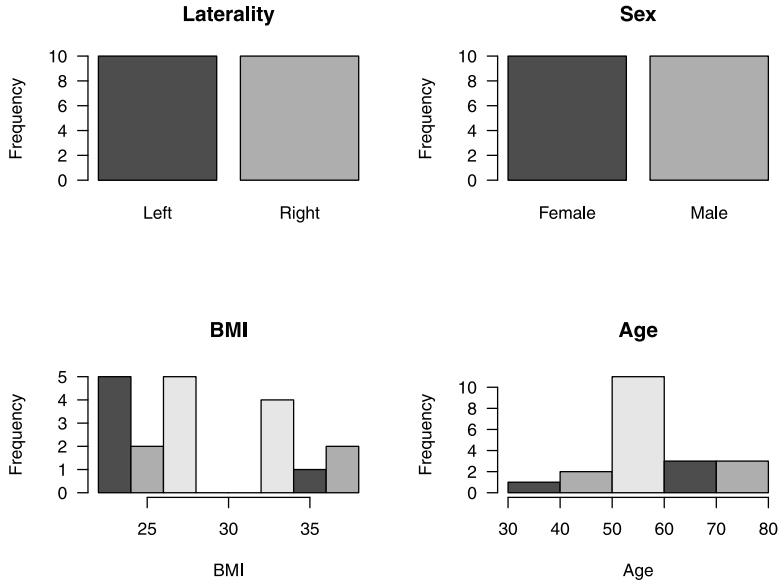


Fig. 3. Distribution of anamnestic data of the Human Glomeruli Dataset.

in this work, a MobileNet V3 [37] is employed for the glomeruli semantic segmentation. The network is designed by using a combination of hardware-aware network architecture search complemented by the NetAdapt algorithm. For the task of semantic segmentation an efficient segmentation decoder called lite reduced atrous spatial pyramid pooling is added to the MobileNet backbone.

4.2. SegNeXt

The SegNeXt [38] model is an innovative approach to semantic segmentation that incorporates an effective encoder–decoder architecture. Unlike traditional CNN models, the SegNeXt uses multi-scale convolutional layers instead of standard convolutional blocks, providing spatial attention at each stage of the encoder with a simple element-wise multiplication. This approach to provide spatial attention has been shown to be more efficient than traditional convolutions and self-attention. The SegNeXt network consists of a convolutional encoder and a decoder that incorporates the decomposition-based Hamburger module, which is used to extract some global context information. In particular, in the decoder, features from different layers are combined and processed through the Hamburger module to provide multiscale context information. As a result, SegNeXt proved to be more accurate than other previous segmentation approaches, including those based on transformers. In this work the small version of the SegNeXt model was employed to reduce the computational burden.

5. Experiments and results

The results of the experiments carried out to validate our glomeruli segmentation approach are described in this section. Specifically, in Section 5.1 the experimental setup is defined, while Section 5.2 describes the metrics used to evaluate the model outputs, and Section 5.3 presents and discuss the obtained results.

5.1. Experimental setup

All experiments were conducted using two segmentation networks, the MobileNet [37] and the SegNeXt [38], pre-trained on the CityScapes dataset [56], on a Linux environment with a single NVIDIA Titan X with 24 GB of memory. Both network architectures were trained using the implementation made available in the MMSegmentation library [57].

Before training the models, image preprocessing was conducted to standardize the size of the glomeruli in the two datasets. Because the images in the datasets were acquired at different resolutions and despite the structural similarities between mouse and human glomeruli, their sizes vary significantly. To optimize the effectiveness of transfer learning, we took a simple approach: we scaled the images in the highest resolution dataset, which is the human dataset, to align the size of the glomeruli with those in the mouse dataset. This adjustment allows for approximate matching of glomeruli size between the two datasets. To be more specific, the average size of a bounding box surrounding a glomerulus in human images is about 385 by 385 pixels, while in mice it is about 154 by 154 pixels. As a result, we applied a scaling factor of 0.4 to the images in the human dataset so that the glomeruli size becomes similar to the mice images.

After the resizing step, all the images were divided into overlapping tiles (50% overlap) having the same size (512×512). A total of 58,818 tiles were extracted from the human dataset and 27,061 from the mice dataset.

The following experimental setup was employed throughout the training of both models. Random flip,[7] random rotation,[8] and photometric distortion[9] were employed to augment the training dataset. The AdamW [58] optimizer, with learning rates of 1×10^{-4} and 6×10^{-5}, was used for training the MobileNet and SegNeXt, respectively. Moreover, a warm-up strategy is employed, starting with a low learning rate (1×10^{-6}) and progressively increasing it for the first 1500 iterations. Cross-entropy loss and a batch size of 18 images were used in all experiments. Additionally, an early stopping approach based on the Intersection over Union (IoU) on the validation set was used to select the best model during training. In the test phase, the images were divided into tiles using the same approach employed in training. The network predictions obtained for all of the tiles are then recombined, and resized to the original image size, to create the final segmentation output. The use of the predictions in the outermost part of the tiles is avoided: in the recombination step, we averaged only the network output probabilities corresponding to the center of each tile. In fact, the network predictions made on the border of the tile may be inaccurate, since if a glomerulus is not fully visible it could be much more difficult to be recognized.

Finally, to demonstrate the effectiveness of cross-species transfer learning, the following experimental setup was devised and tested with both network architectures. First, the network was trained on the mice dataset and on the human dataset, independently. Then, the model trained on the mice data was fine-tuned using the human training set. Fine-tuning is a common practice in machine learning that involves adapting a pre-trained model to a specific task or dataset to enhance its performance. In the context of transfer learning, fine-tuning has been proven to be effective in various computer vision applications [59]. With fine-tuning, the weights learned by the model on one domain (in our case, the mice dataset) are adjusted to capture task-specific patterns in the target domain (the human dataset). This process allows the model to specialize and improve its performance on the target task. Depending on the availability of data, it is possible to fine-tune all the model's weights or only a subset of them. In our application, we choose to apply the cross-species transfer learning with two approaches: by fine-tuning all the weights of the segmentation network and by fine-tuning only the decoder. This approach enables the model to leverage the knowledge gained from the mice dataset and adapt it to the human dataset, leading to improved segmentation performance. To demonstrate the effectiveness of the proposed approach, we compare the results obtained with and without the application of the fine-tuning procedure. To evaluate the stability and repeatability of the results obtained through the fine-tuning approach, we conducted a four-folds cross-validation procedure. Specifically, we randomly selected three images from the human dataset four times, each time designating them as the validation set. This process allowed us to retrain the network with different training–validation splits and, for each iteration, we evaluated the model's performance on the public test set. To statistically evaluate whether the results obtained with networks trained exclusively on human data are not in line with the distribution derived from the four cross-validation results, we used a one-sample t-test.

5.2. Evaluation metrics

To assess segmentation performance, we utilized two widely used metrics: Jaccard, also referred to as IoU [60], and the Dice index [61]. Given two generic sets A and B, the Jaccard and Dice indices are defined in Eqs. (1) and (2):

$$Jaccard\ (IoU) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

[7] Horizontal and vertical.

[8] 90°, 180° and 270° degrees.

[9] Delta brightness equal to 32, contrast range from 0.5 to 1.5, saturation range from 0.5 to 1.5 and delta hue of 18.

Table 1
Results on the validation set and on the test set of the mice dataset obtained by the two segmentation models.

Models	Validation set		Test set	
	MIoU	Dice	MIoU	Dice
MobileNet	83.96%	90.27%	83.03%	89.46%
SegNeXt	86.27%	92.01%	85.15%	91.12%

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

Both metrics are commonly used to evaluate the output of a semantic segmentation model, and for this reason, we have decided to use both in our study. When used for semantic segmentation performances assessments, the two sets A and B are respectively the set of the pixels in the ground truth mask and of the pixels in the segmentation mask produced by the deep neural network.

Furthermore, because our goal is to accurately count the glomeruli, we employ a straightforward metric that assesses the difference between the number of glomeruli annotated in the ground truth and those identified by the network. To determine the glomeruli count from a segmentation mask, we leverage the concept of connected components, counting the distinct regions within the mask. The $\Delta_{Glom.}$ metric, which is defined as the absolute difference between the number of glomeruli present in the target mask (GT) and the number predicted by the network ($Pred$), was then defined to evaluate the performance of the models (see Eq. (3)).

$$\Delta_{Glom.} = |num\_glomeruli(GT) - num\_glomeruli(Pred)| \tag{3}$$

5.3. Results

In the following subsections we present the results obtained following the proposed experimental setup. In Section 5.3.1 the results of the glomeruli segmentation process for mice are discussed. Meanwhile, in Section 5.3.2, we compare the segmentation results on human glomeruli with and without the use of the cross-species pre-training.

5.3.1. Mice glomeruli segmentation

Following the experimental setup described in the previous section, we first trained the two network architectures to segment mice glomeruli. In Table 1 the results of the models on the validation set and on the test set of the mice dataset are reported. Moreover, a qualitative evaluation of the segmentation outputs obtained on the test set is shown in Fig. 4.

As we can observe from these results both models are able to segment mice glomeruli quite accurately.

5.3.2. Human glomeruli segmentation

The main goal of this study is to investigate whether the features learned on mice glomeruli can be used to enhance the generalization of a model on human images. Therefore we trained and compared the results obtained by the two segmentation networks with and without the pre-training on mice glomeruli.

Evaluation on validation and test set.

In Table 2 the performances on the validation set of the HuBMAP glomeruli dataset are reported.

As it can be observed, employing the cross-species transfer learning approach allows to obtain significant improvements in the performance on the validation set for both the MobileNet and SegNeXt models. Moreover, it is noteworthy that, by employing fine-tuning, there is a reduction in the average difference between the actual and the predicted number of glomeruli (Avg. $\Delta_{Glom.}$) for both MobileNet and SegNeXt. Specifically, the MobileNet and SegNeXt models exhibit an average improvement in the $\Delta_{Glom.}$ of approximately 10 and 25 glomeruli per

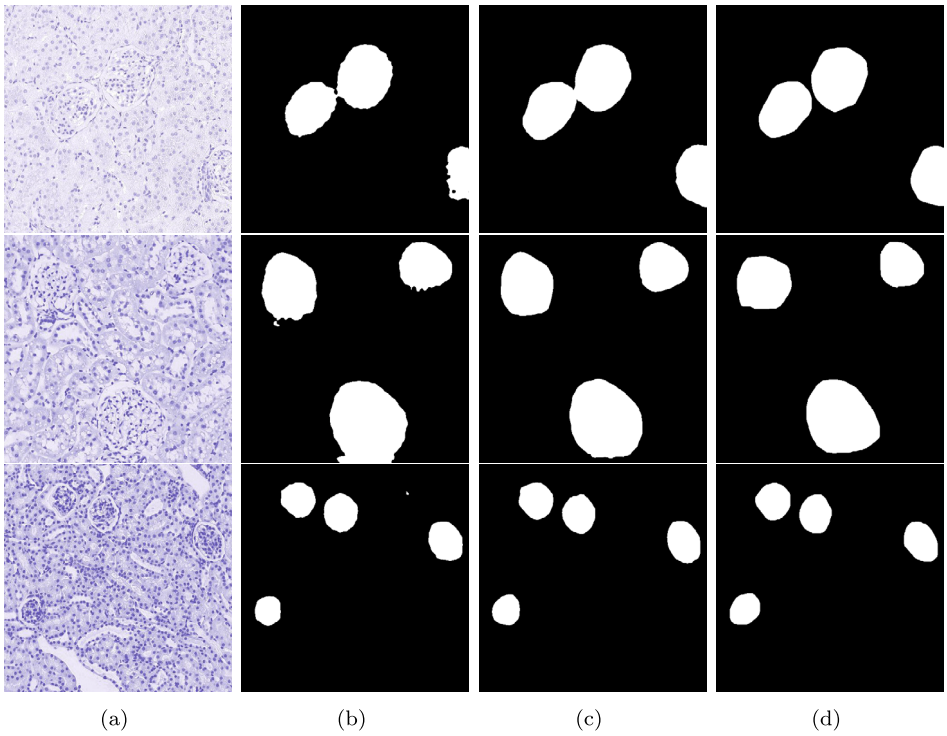


Fig. 4. Segmentation results on the test set of the mice dataset. In (a) the original image crops, in (b) and (c) the segmentations obtained with the MobileNet and the SegNeXt, respectively, and in (d) the ground truth annotations.

Table 2
Segmentation results on the validation set of the HuBMAP dataset with the MobileNet and the SegNeXt.

Training	MobileNet			SegNeXt		
	MIoU	Dice	Avg $\Delta_{Glom.}$	MIoU	Dice	Avg $\Delta_{Glom.}$
Human	90.04%	94.49%	24.66	91.44%	95.33%	30.33
Mice+Human	91.83%	95.56%	13.0	92.97%	96.23%	5.33

Table 3
Segmentation results on the test set of the HuBMAP dataset with the MobileNet and the SegNeXt.

Training	MobileNet Dice	SegNeXt Dice
Human	82.67%	86.85%
Mice+Human	85.61%	90.43%

image, respectively. Interestingly, despite SegNeXt being a considerably more powerful model than MobileNet, MobileNet outperforms SegNeXt in terms of Avg. $\Delta_{Glom.}$ when fine-tuning is not employed. This suggests that although the SegNeXt model yields higher values for MIoU and Dice, the segmentation maps it produces may be less precise in terms of connected components (f.i. fusing or splitting glomeruli). This observation might be attributed to the fact that a network with fewer parameters possesses better generalization capabilities when trained with limited data. Indeed, when cross-species transfer learning is applied, and the mice images are employed to enlarge the training images, the results for the SegNeXt model show a substantial improvement.

The models were also evaluated on the test set of the HuBMAP dataset. In particular, to compute the score on this set of data, the model output segmentation masks were submitted on the official evaluation server of the contest[10] and the results are presented in Table 3.

Additionally, Figs. 5 and 6 show a qualitative evaluation of the segmentation obtained by the models on randomly selected crops from the validation set and from the test set, respectively.

The results show that the pre-training on mice data allows to increase the performance with both network architectures. In fact, both models achieve about a 3% improvement in the Dice index when pre-trained on mice data.

[10] https://www.kaggle.com/competitions/hubmap-kidney-segmentation/submissions.

Hence, our approach is useful not only for models with a very large number of parameters like the SegNeXt (13.9M parameters) but also for lighter models like the MobileNet V3 (3.3M parameters). As expected the SegNeXt, which has a larger number of parameters than the MobileNet, produces the best results with a Dice score of 90.43% on the public test set.

Moreover, qualitatively, Figs. 5 and 6 provide an illustrative example of the segmentation results obtained on the validation set and on the test set, respectively. In Fig. 5, the second row shows an image with a glomerulus with atypical shapes and colors, which poses a challenge for both networks when trained only on the human dataset. Similarly, in Fig. 6, it can be observed that the glomerulus in the third row, with an unusual shape, is correctly recognized only by the SegNeXt model after fine-tuning. Indeed, if the cross-species pre-training is employed the results are significantly improved for both networks.
Furthermore, it is interesting to be observed that the SegNext is able to overcome the MobileNet when the pre-training is employed. This suggests that the use of data from different species can be an effective type of data augmentation allowing to improve the generalization of more complex models. Overall, these observations confirm the potential of cross-species pre-training to provide more accurate results and better generalization.

Evaluation of the fine-tuning procedure freezing the initial network layers.

We conducted further experiments aimed at evaluating the potential benefits of specializing the final layers by freezing earlier ones. Specifically, for both network architectures pre-trained on mice data, we froze

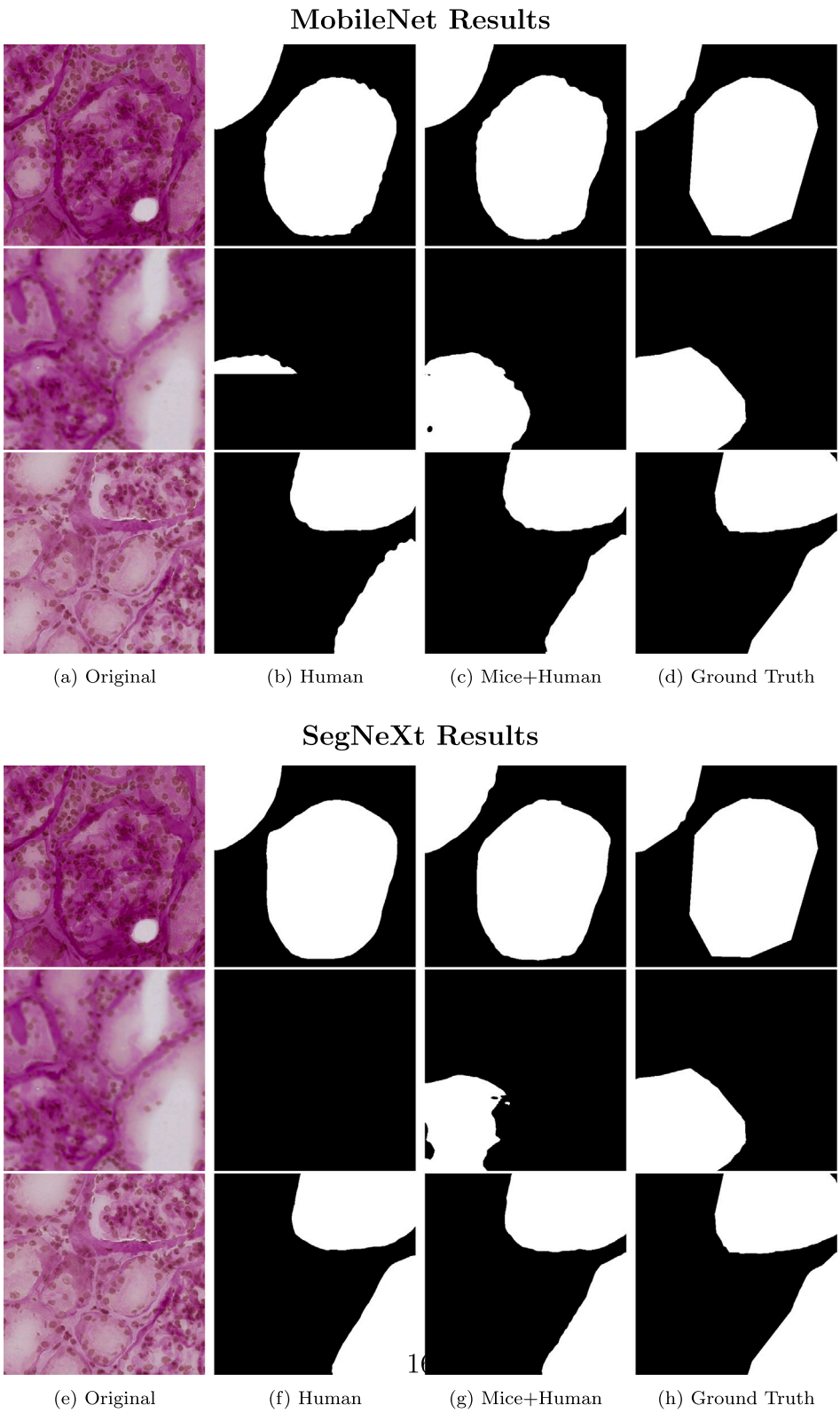


Fig. 5. Segmentation results from the two models, MobileNet (top) and SegNeXt (bottom), on the validation set of the HuBMAP dataset. In (a and e) the original images, in (b and f) the segmentation results from the model trained solely on human data, in (c and g) the segmentation results obtained through cross-species transfer learning, and in (d and h) the ground truths.

the encoder layers and trained the decoder exclusively. However, the results obtained on both the validation and test sets were worse to those obtained without freezing the initial network layers. On the validation set, we recorded a Dice Score of 92.83% and a mIoU of 87.42% with SegNeXt, while the MobileNet produced a Dice Score of 86.22% and a mIoU of 78.72%. This difference persisted when evaluating the models on the test set, where SegNeXt produced a Dice Score of 81.72% and MobileNet produced a Dice Score of 75.19%. This might

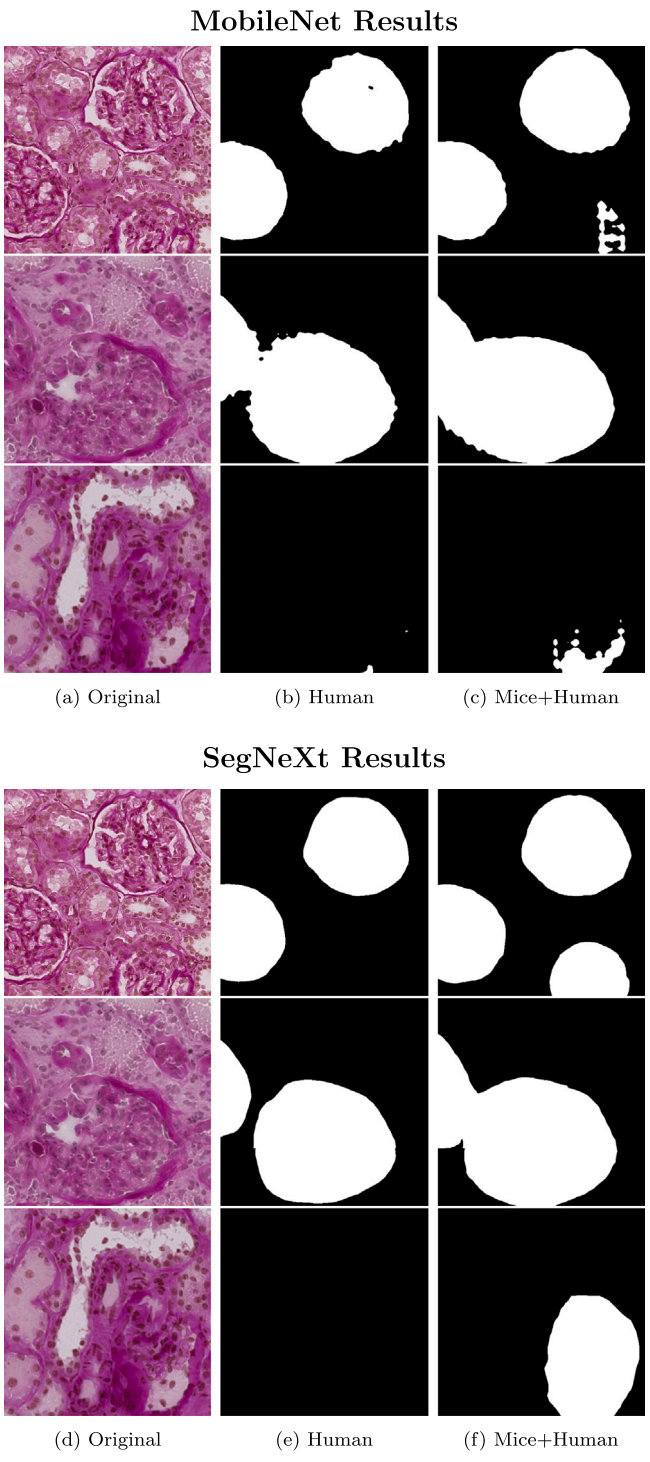


Fig. 6. Segmentation results from the two models, MobileNet (top) and SegNeXt (bottom), on the test set of the HuBMAP dataset. In (a and d) the original images, in (b and e) the segmentation results from the model trained solely on human data, and in (c and f) the segmentation results obtained through cross-species transfer learning.

confirm that the features learned from mouse data are indeed valuable, allowing us to achieve satisfactory segmentation results on human data. However, although there are shared features between mice and humans glomeruli, the images show strongly different characteristics.
For this reason, it is not surprising that fine-tuning the encoder also helps refine the feature extractor specific to human glomeruli, improving the overall segmentation performance.

Table 4
Results of the four-fold cross-validation.

Models	MIoU			Dice		
	Mean	Std	*pvalue*	Mean	Std	*pvalue*
MobileNet	91.86%	0.46	0.004	95.57%	0.28	0.004
SegNeXt	92.98%	0.51	0.009	96.22%	0.31	0.010

Table 5
Results on the test set of the model trained in the four-folds.

Models	Dice		
	Mean	Std	*pvalue* MIoU
MobileNet	85.46%	0.48	0.001
SegNeXt	89.17	1.18	0.029

Repeatability evaluation.

To assess the repeatability of the results obtained through cross-species pre-training, we conducted a four-folds cross-validation. Table 4 presents the mean and standard deviation of both mIoU and Dice scores across the four folds. Additionally, we include the results of one-sample t-tests comparing the Dice and mIoU scores obtained in the four runs with the same values achieved training the model without the cross-species pre-training.

As we can observe, the results show statistical consistency with those detailed in the previous paragraph, where we used a single training–validation split. Indeed, both SegNeXt and MobileNet show low standard deviation. Furthermore, the p-value consistently falls below 0.05, indicating that the mean obtained in the absence of cross-species transfer learning is not in line with the distribution generated from the means obtained with the four-fold cross-validation. This further highlights that the cross-species transfer learning approach outperforms the direct training on human data. Given that the t-test assumes that the analyzed samples follow a Gaussian distribution, we conducted a Shapiro–Wilk test to assess this assumption. The resulting p-values consistently exceeded 0.05, suggesting that the data could indeed be Gaussian. Furthermore, we evaluated the models trained with the four training–validation splits on the public test set. In Table 5, we reported the mean and standard deviation of the Dice scores, along with their respective p-values, obtained with the four runs.

The results are similar also for the test set, where the p-values, consistently below 0.05, provide further confirmation of the effectiveness of the proposed cross-species transfer learning approach.

6. Conclusion

In this study, we applied two DL architectures, SegNeXt and MobileNet, to the segmentation of glomeruli in human kidneys. The core idea of the proposed method is the use of a cross-species transfer learning approach, which consists in pre-training the models on mice glomeruli images and in fine-tuning it on human data. The results were promising, with the pre-training allowing for improved segmentation performance for both SegNeXt and MobileNet. Moreover, the cross-species pre-training proved to be effective in overcoming the shortage of annotated human glomeruli images and increasing the generalization capabilities of the models. Indeed, without the pre-training, the networks were often unable to correctly recognize glomeruli with unusual color or shape. This work can be the base for more accurate and efficient analysis, reducing the possibility of human error and saving time. Furthermore, the successful results obtained using MobileNet hold great potential for developing a DL-based glomeruli segmentation tool that could be used in mobile or embedded environments (i.e. electron microscopes). As a future research direction, it would be interesting to explore the extension of the segmentation models to different staining procedures. Additionally, distinguishing between sclerotized and non-sclerotized glomeruli could be important for the diagnosis of certain kidney diseases.

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.

Data availability

All the data used is publicly available.

References

[1] C. Feng, F. Liu, Artificial intelligence in renal pathology: Current status and future, Biomol. Biomed. (2023).

[2] J.J. Hogan, M. Mocanu, J.S. Berns, The native kidney biopsy: update and evidence for best practice, Clin. J. Am. Soc. Nephrol. 11 (2) (2016) 354–362.

[3] B.J. Nankivell, S.M. Gruenewald, R. Allen, J.R. Chapman, Predicting glomerular filtration rate after kidney transplantation., Transplantation 59 (12) (1995) 1683–1689.

[4] R.W. Major, Clinical assessment of kidney function and prognosis in adults, Medicine (2023).

[5] M. Windpessl, M. Kostopoulou, R. Conway, I.B. Mentese, A. Bruchfeld, M.J. Soler, M. Sester, A. Kronbichler, Preventing infections in immunocompromised patients with kidney diseases: vaccines and antimicrobial prophylaxis, Nephrol. Dial. Transplant. (2023) gfad080.

[6] S. Abramyan, M. Hanlon, Kidney transplantation, 2021.

[7] L.A. Stevens, J. Coresh, T. Greene, A.S. Levey, Assessing kidney function—measured and estimated glomerular filtration rate, N. Engl. J. Med. 354 (23) (2006) 2473–2483.

[8] J. Karpinski, G. Lajoie, D. Cattran, S. Fenton, J. Zaltzman, C. Cardella, E. Cole, Outcome of kidney transplantation from high-risk donors is determined by both structure and function, Transplantation 67 (8) (1999) 1162–1167.

[9] J. Taylor, Renal system 4: causes, diagnosis and treatment of chronic kidney disease, Nursing Times (2023).

[10] A. Young, S.J. Kim, A.X. Garg, A. Huang, G. Knoll, G.R. Prasad, D. Treleaven, C.E. Lok, D.N.O.R.D. Network, J. Arnold, et al., Living kidney donor estimated glomerular filtration rate and recipient graft survival, Nephrol. Dial. Transplant. 29 (1) (2014) 188–195.

[11] V.D. D'Agati, F.J. Kaskel, R.J. Falk, Focal segmental glomerulosclerosis, N. Engl. J. Med. 365 (25) (2011) 2398–2411.

[12] G.M. Dimitri, P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, F. Scarselli, A. Zacchi, G. Garosi, T. Marcuzzo, S.A. Tripodi, Deep learning approaches for the segmentation of glomeruli in kidney histopathological images, Mathematics 10 (11) (2022) 1934.

[13] S. Kannan, L.A. Morgan, B. Liang, M.G. Cheung, C.Q. Lin, D. Mun, R.G. Nader, M.E. Belghasem, J.M. Henderson, J.M. Francis, et al., Segmentation of glomeruli within trichrome images using deep learning, Kidney Int. Rep. 4 (7) (2019) 955–962.

[14] A. Akbari, J. Grimshaw, D. Stacey, W. Hogg, T. Ramsay, M. Cheng-Fitzpatrick, P. Magner, R. Bell, J. Karpinski, Change in appropriate referrals to nephrologists after the introduction of automatic reporting of the estimated glomerular filtration rate, CMAJ 184 (5) (2012) E269–E276.

[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[17] O. Spiga, V. Cicaloni, G.M. Dimitri, F. Pettini, D. Braconi, A. Bernini, A. Santucci, Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease, Brief. Bioinform. 22 (5) (2021) bbaa434.

[18] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M.M. Hoffman, et al., Opportunities and obstacles for deep learning in biology and medicine, J. R. Soc. Interface 15 (141) (2018) 20170387.

[19] P. Bongini, N. Pancino, G.M. Dimitri, M. Bianchini, F. Scarselli, P. Lio, Modular multi–source prediction of drug side–effects with drugnn, IEEE/ACM Trans. Comput. Biol. Bioinform. (2022).

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[21] D.W. Otter, J.R. Medina, J.K. Kalita, A survey of the usages of deep learning for natural language processing, IEEE Trans. Neural Netw. Learn. Syst. 32 (2) (2020) 604–624.

[22] P. Andreini, G. Ciano, S. Bonechi, C. Graziani, V. Lachi, A. Mecocci, A. Sodi, F. Scarselli, M. Bianchini, A two-stage GAN for high-resolution retinal image generation and segmentation, Electronics 11 (1) (2021) 60.

[23] S. Bonechi, Isic_Wsm: Generating weak segmentation maps for the ISIC archive, Neurocomputing 523 (2023) 69–80.

[24] S. Bonechi, P. Andreini, A. Mecocci, N. Giannelli, F. Scarselli, E. Neri, M. Bianchini, G.M. Dimitri, Segmentation of aorta 3D CT images based on 2D convolutional neural networks, Electronics 10 (20) (2021) 2559.

[25] A. Rossi, G. Vannuccini, P. Andreini, S. Bonechi, G. Giacomini, F. Scarselli, M. Bianchini, Analysis of brain NMR images for age estimation with deep learning, Procedia Comput. Sci. 159 (2019) 981–989.

[26] G.M. Dimitri, S. Spasov, A. Duggento, L. Passamonti, P. Lió, N. Toschi, Multimodal and multicontrast image fusion via deep generative models, Inf. Fusion 88 (2022) 146–160.

[27] D. Meconcelli, S. Bonechi, G.M. Dimitri, Deep learning approaches for mice glomeruli segmentation, in: ESANN 2022 Proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2022.

[28] M. Stritt, A.K. Stalder, E. Vezzali, Orbit image analysis: an open-source whole slide image analysis tool, PLoS Comput. Biol. 16 (2) (2020) e1007313.

[29] M.F. Lyon, E. Hulse, An inherited kidney disease of mice resembling human nephronophthisis., J. Med. Genet. 8 (1) (1971) 41.

[30] N.O. Lindström, J.A. McMahon, J. Guo, T. Tran, Q. Guo, E. Rutledge, R.K. Parvez, G. Saribekyan, R.E. Schuler, C. Liao, et al., Conserved and divergent features of human and mouse kidney organogenesis, J. Am. Soc. Nephrol. 29 (3) (2018) 785–805.

[31] M. Abdalla, M. Abdalla, F.S. Siddiqi, L. Geldenhuys, S.N. Batchu, M.F. Tolosa, D.A. Yuen, C.C. Dos Santos, A. Advani, A common glomerular transcriptomic signature distinguishes diabetic kidney disease from other kidney diseases in humans and mice, Curr. Res. Transl. Med. 68 (4) (2020) 225–236.

[32] M.A. Brehm, A.C. Powers, L.D. Shultz, D.L. Greiner, Advancing animal models of human type 1 diabetes by engraftment of functional human tissues in immunodeficient mice, Cold Spring Harb. Perspect. Med. 2 (5) (2012) a007757.

[33] M.S. Burhans, D.K. Hagman, J.N. Kuzma, K.A. Schmidt, M. Kratz, Contribution of adipose tissue inflammation to the development of type 2 diabetes mellitus, Compr. Physiol. 9 (1) (2018) 1.

[34] Y. Sato, P. Boor, S. Fukuma, B.M. Klinkhammer, H. Haga, O. Ogawa, J. Floege, M. Yanagita, Developmental stages of tertiary lymphoid tissue reflect local injury and inflammation in mouse and human kidneys, Kidney Int. 98 (2) (2020) 448–463.

[35] D.D. Shapiro, M. Virumbrales-Muñoz, D.J. Beebe, E.J. Abel, Models of renal cell carcinoma used to investigate molecular mechanisms and develop new therapeutics, Front. Oncol. 12 (2022).

[36] J.M. Overstreet, C.C. Gifford, J. Tang, P.J. Higgins, R. Samarakoon, Emerging role of tumor suppressor p53 in acute and chronic kidney diseases, Cell. Mol. Life Sci. 79 (9) (2022) 474.

[37] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q.V. Le, H. Adam, Searching for MobileNetV3, in: The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324, http://dx.doi.org/10.1109/ICCV.2019.00140.

[38] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, S.-M. Hu, Segnext: Rethinking convolutional attention design for semantic segmentation, 2022, arXiv preprint arXiv:2209.08575.

[39] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[40] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv:1706.05587.

[41] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[42] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 11–19.

[43] F. Lateef, Y. Ruichek, Survey on semantic segmentation using deep learning techniques, Neurocomputing 338 (2019) 321–348.

[44] J. Van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic, Nat. Med. 27 (5) (2021) 775–784.

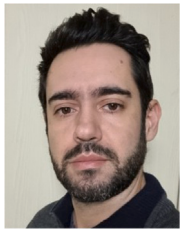
[45] G. Bueno, M.M. Fernandez-Carrobles, L. Gonzalez-Lopez, O. Deniz, Glomerulosclerosis identification in whole slide images using semantic segmentation, Comput. Methods Programs Biomed. 184 (2020) 105273.

[46] N. Altini, G.D. Cascarano, A. Brunetti, I. De Feudis, D. Buongiorno, M. Rossini, F. Pesce, L. Gesualdo, V. Bevilacqua, A deep learning instance segmentation approach for global glomerulosclerosis assessment in donor kidney biopsies, Electronics 9 (11) (2020) 1768.

[47] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.322.

[48] C.P. Jayapandian, Y. Chen, A.R. Janowczyk, M.B. Palmer, C.A. Cassol, M. Sekulic, J.B. Hodgin, J. Zee, S.M. Hewitt, J. O'Toole, et al., Development and evaluation of deep learning–based segmentation of histologic structures in the kidney cortex with multiple histologic stains, Kidney Int. 99 (1) (2021) 86–101.

[49] H.-C. Lee, A.F. Aqil, Combination of transfer learning methods for kidney glomeruli image classification, Appl. Sci. 12 (3) (2022) 1040.
[50] R. Statkevych, Y. Gordienko, S. Stirenko, Improving U-net kidney glomerulus segmentation with fine-tuning, dataset randomization and augmentations, in: Advances in Computer Science for Engineering and Education, Springer, 2022, pp. 488–498.
[51] F.N. Saikia, Y. Iwahori, T. Suzuki, M. Bhuyan, A. Wang, B. Kijsirikul, MLP-unet: Glomerulus segmentation, IEEE Access (2023).
[52] J. Silva, L. Souza, P. Chagas, R. Calumby, B. Souza, I. Pontes, A. Duarte, N. Pinheiro, W. Santos, L. Oliveira, Boundary-aware glomerulus segmentation: Toward one-to-many stain generalization, Comput. Med. Imaging Graph. 100 (2022) 102104.
[53] P. Sarder, B. Ginley, J.E. Tomaszewski, Automated renal histopathology: digital extraction and quantification of renal pathology, in: Medical Imaging 2016: Digital Pathology, vol. 9791, SPIE, 2016, pp. 112–123.
[54] 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 (9) (2019) 1786–1796.
[55] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[56] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[57] M. Contributors, MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark, 2020, https://github.com/open-mmlab/mmsegmentation.
[58] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2017.
[59] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021) 104115.
[60] A.H. Murphy, The finley affair: A signal event in the history of forecast verification, Weather Forecast. 11 (1) (1996) 3–20.
[61] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.



Paolo Andreini is a researcher at the University of Siena (Italy), has a degree in computer science and a Ph.D. in Information Engineering and Science. His Ph.D. thesis focused on proposing new generative models that can be trained from very limited datasets and can thus be used in the absence of large sets of data. His research activity focuses on machine learning and computer vision. The main research topics are object detection, semantic and instance segmentation, semi-supervised and deep learning. From a practical point of view, he worked on a variety of topics including text detection and recognition (in natural and cluttered images), agar plate analysis, retinal vessel segmentation, skin lesion classification and brain age estimation from MR images.

Simone Bonechi graduated in Computer Science at the Department of Information Engineering and Mathematical Sciences of the University of Siena (2014) where he then obtained his PhD. In 2018 he spent a period as a visiting Ph.D. at the University of Copenhagen (2018). After two years of PostDoc, first at the University of Tuscia, and then at the University of Pisa, he is now a researcher at the Department of Social, Political and Cognitive Sciences of the University of Siena. His research activity is focused on Deep Learning and Artificial Intelligence, with particular reference to computer vision, image processing and image generation. He is the author of over 25 publications in international journals and conference proceedings. He also carries out editorial activities as Associated Editor for the journal Neurocomputing and he is Guest Editor for the Special Issue "Mathematical Modelling and Machine Learning Methods for Bioinformatics and Data Science Applications II" for the journal Mathematics.

Giovanna Maria Dimitri is a researcher at the DIISM, University of Siena, Italy. Previously she obtained the Ph.D. at the University of Cambridge (UK), supervised by Prof. Pietro Liò, with the dissertation: "Multilayer network methodologies for brain data analysis and modelling". She graduated in July 2015 in the MPhil in Advanced Computer Science at the University of Cambridge, with distinction. She is a life member of Clare Hall college, University of Cambridge. Previously she received her master and bachelor thesis (both 110/110 cum laude) in Computer and Automation Engineering at the University of Siena (Italy), supervised by Prof. Michelangelo Diligenti. She is lecturing the Business Intelligence course for the master in Engineering Management (DIISM, University of Siena) since A.Y. 2019/2020. She is associate editor for Neurocomputing. Her research interests mainly concerns developing deep learning and machine learning models for computer vision and bioinformatics.