Recruiting the Best Teacher Modality: A Customized Knowledge Distillation Method for if Based Nephropathy Diagnosis

Ning Dai[1], Lai Jiang[1(✉)], Yibing Fu[1], Sai Pan[2], Mai Xu[1], Xin Deng[3], Pu Chen[2], and Xiangmei Chen[2]

[1] School of Electronic and Information Engineering, Beihang University, Beijing, China
jianglai.china@buaa.edu.cn

[2] National Clinical Research Center for Kidney Diseases, State Key Laboratory of Kidney Diseases, Institute of Nephrology of Chinese PLA, Department of Nephrology, General Hospital of Chinese PLA, Medical School of Chinese PLA, Beijing, China

[3] School of Cyber Science and Technology, Beihang University, Beijing, China

Abstract. The joint use of multiple imaging modalities for medical image has been widely studied in recent years. The fusion of information from different modalities has demonstrated the performance improvement for some medical tasks. For nephropathy diagnosis, immunofluorescence (IF) is one of the most widely-used medical image due to its ease of acquisition with low cost, which is also an advanced multi-modality technique. However, the existing methods mainly integrate information from diverse sources by averaging or combining them, failing to exploit multi-modality knowledge in details. In this paper, we observe that the 7 modalities of IF images have different impact on different nephropathy categories. Accordingly, we propose a knowledge distillation framework to transfer knowledge from the trained single-modality teacher networks to a multi-modality student network. On top of this, given a input IF sequence, a recruitment module is developed to dynamically assign weights to teacher models and optimize the performance of student model. By applying on several different architectures, the extensive experimental results verify the effectiveness of our method for nephropathy diagnosis.

Keywords: Nephropathy diagnosis · IF image · Knowledge distillation

This work was supported by NSFC under Grant 62250001, and Alibaba Innovative Reaserch.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_51.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 526–536, 2023.
https://doi.org/10.1007/978-3-031-43904-9_51

1 Introduction

Nephropathy is a progressive and incurable disease with high mortality, occurring commonly in the general adult population, with a world-wide prevalence of 10% [11]. Therefore, early detection and treatment is of pivotal importance, as it can prevent the death or inevitable renal failure that requires renal dialysis or replacement therapy. Due to the low cost and sensitivity for certain lesion [20], immunofluorescence (IF) images have been increasingly used in the diagnostic process of nephropathy. Most recently, benefiting from the development of deep learning, a couple of deep neural networks (DNNs) have been developed for nephropathy related tasks on IF images [7,8,13,18]. For instance, Ligabue *et al.* [8] proposed a residual convolutional neural network (CNN) for IF nephropathy reporting. Similarly, in [18], a DNN-based model with a pre-segmentation module and a classification module was introduced to automatically detect the different glomerulus in IF images. Kitamura *et al.* [7] designed a CNN structure for diabetic nephropathy (DN) diagnosis on IF images, and further visualized where the CNN focused on for diagnosis.

There exist only a few IF image based DNN methods, probably because the properties of IF images are complicated and have not been fully exploited. Different from the natural and other medical images, a IF sequence usually include multiple modalities from different types of fluorescent [2]. On the other hand, the collected modalities of a IF sequence are usually incomplete, due to the medical reasons and acquiring processes. This leads to various modality combination in a IF dataset, therefore significantly reducing the learning efficiency of a DNN. More importantly, the correlation between different IF modalities and nephropathy categories is complicated. For instance, as shown in Fig. 1, the experiments in this paper find that anti-neutrophil cytoplasmic antibodies (ANCA) disease is strongly related to the modalities of Immunoglobulin G (IgG) and Immunoglobulin A (IgA). Meanwhile, other modalities like Complement 3

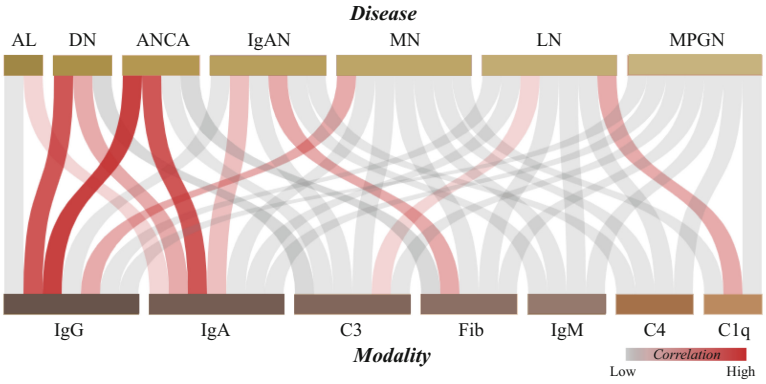

Fig. 1. The importance of a individual modality contributing to the diagnosis on each nephropathy. The detailed calculation is introduced in the data analysis section.

(C3) and Fibronectin (Fib) sometimes mislead the DNN and thus degrade the ANCA diagnosis performance. Unfortunately, all above DNN methods assume different modalities have the equal effect on the diagnosis task, neglecting the medical prior of the importance of individual modality.

To address above issue, this paper proposes a novel customized multi-teacher knowledge distillation framework for nephropathy diagnosis on IF images. Specifically, we establish a large-scale IF image dataset including 7 types of nephropathy from 1,582 patients. By mining our dataset, we conduct experiments to explore the importance of a individual modality contributing to each nephropathy, as the empirical medical prior (see Fig. 1). Then, we develop a multi-teacher knowledge distillation framework, in which the knowledge is transferred from the teacher networks trained by individual modalities. Different from the traditional knowledge distillation [5,9], we propose a customized framework with a recruitment module, which learns to select the "best" teacher networks based on the medical priors. Benefiting from this, the student network can effectively learns from the individual modalities, thus achieving better overall performance for the clinical IF sequence with incomplete modalities. We show the effectiveness of the proposed method over our dataset and another external dataset for nephropathy diagnosis. In summary, the main contributions of this paper are three-fold.

- We establish a large-scale IF dataset containing 7 nephropathy categories and 1,582 IF sequences, with the data analysis for the importance of a individual modality contributing to the diagnosis on each nephropathy category.
- We propose a new customized knowledge distillation framework for nephropathy diagnosis, which transfers the knowledge from individual modalities for improving the performance over IF sequence with multiple modalities.
- We develop a recruitment module in our knowledge distillation method, which learns to select the "best" teacher modalities based on the medical priors of input IF sequence and its nephropathy.

2 Dataset Establishment and Analysis

2.1 Dataset Establishment

We first propose a large-scale dataset with 1,582 IF sequences and 6,381 images as our main dataset, including 7 categories of nephropathy, i.e., membranous nephropathy (MN), IgA nephropathy (IgAN), lupus nephropathy (LN), DN, ANCA, membranoproliferative glomerulonephritis (MPGN) and Amyloidosis (AL). Each IF sequence has at most 7 modalities, including IgG, IgA, Immunoglobulin M (IgM), C3, Complement 4 (C4), Complement 1q (C1q) and Fib. The IF images were acquired from the kidney specimens by a fluorescence microscope following a standardized protocol. Then, each IF sequence was diagnosed by professional nephrologists, together with other medical information, such as medical history and examination results of patient. Besides, with similar processes, we further establish an external dataset collected from another

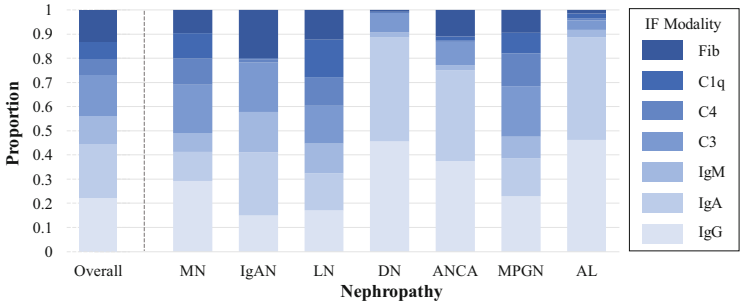

Fig. 2. The proportions of each IF modality for the 7 categories of nephropathy.

hospital including 69 IF sequences with 348 images. Note that our work is conducted according to the Declaration of Helsinki. Compared with the existing IF image datasets [7,8], our dataset collects most patients including most categories of nephropathy. In the experiments, our main dataset is randomly divided into training and test sets at a ratio of 4 to 1, while the external set is only used for test.

2.2 Dataset Analysis

Based on our main dataset, we further conduct data analysis to obtain the following findings about the relationship between different modalities and nephropathy.

Finding 1: *The proportions of each IF modality vary greatly in different nephropathy.*

Analysis: As introduced above, the collected modalities of a IF sequence are usually incomplete. This is partially due to the importance of each IF modality for the specific nephropathy, since the patient may not have the anti-body of the useless fluorescent. Therefore, we count the proportions of 7 IF modalities on each nephropathy over 1,582 IF sequences in our main set. As shown in Fig. 2, there exists a significant inconsistency of the modality proportions between different nephropathy. For instance, IgAN does not have the modality of C1q, while LN has almost equal proportions for all IF modalities. Besides, the proportions of a certain IF modality vary a lot in different nephropathy. For example, the proportions of IgM are 10.9%, 1.4%, 11.9%, 0, 0.7%, 13.5%, 0.7% in 7 modalities, respectively. The above analysis completes the analysis of *Finding 1.*

Finding 2: *For certain combinations, the single-modality IF image achieves better diagnosis accuracy than multi-modality IF sequences over DNN models.*

Analysis: To explore the impact of each modality in DNN models, we conduct experiments to evaluate the effectiveness of single IF modality for diagnosing each nephropathy. Specifically, we compare the nephropathy diagnosis accuracy of the same DNN model, when trained and tested over multi-modality IF sequences versus single-modality IF images. First, two widely-used classification models (ResNet-18 [4] and ECANet [15]) are implemented for 7-class

nephropathy diagnosis. Then, we construct 7 dataset pairs from our main set, according to each IF modality. For each data pair, the DNN models are trained and tested with multi-modality and single-modality, respectively, the diagnosis accuracy of which is recorded as Acc_m and Acc_s. Subsequently, these two kinds of accuracy are compared by calculating the error weight E as follows,

$$E = \begin{cases} (\mathrm{Acc}_s - \mathrm{Acc}_m)/\mathrm{Acc}_m, & \mathrm{Acc}_s < \mathrm{Acc}_m \\ (\mathrm{Acc}_s - \mathrm{Acc}_m)/(1 - \mathrm{Acc}_m), & \mathrm{Acc}_s > \mathrm{Acc}_m \end{cases} \tag{1}$$

Thus, the higher error weight indicates that the single-modality can achieve more accurate diagnosis compared with using all modalities. Table 1 tabulates the error weights between each pair of IF modality and nephropathy, over ResNet-18 and ECANet. As shown, for certain combinations, such as IgG to ANCA, IgA to DN, and Fib to IgAN, the single-modality can even achieve better performance than using multi-modality IF sequences. Besides, the phenomenon is consistent over multiple DNN models. This implies that there exists a correlation for each single modality for contributing to the diagnosis on each nephropathy.

Table 1. The error weight results over ResNet-18/ECANet. Note that the error weights are marked in **bold** when consistently positive over two DNN models.

		IF Modality						
		IgG	IgA	IgM	C3	C4	C1q	Fib
Nephropathy	MN	**0.42**/**0.33**	-0.42/-0.31	-0.47/-0.53	-0.27/-0.34	-0.30/-0.37	-0.08/-0.26	-0.85/-0.80
	IgAN	-0.03/-0.03	**0.29**/**0.17**	-0.20/-0.11	-0.08/-0.09	-0.76/-0.81	0/0	**0.50**/**0.25**
	LN	-0.26/-0.22	-0.35/-0.28	-0.18/-0.15	**0.08**/**0.17**	-0.18/-0.14	**0.27**/**0.38**	-0.61/-0.11
	DN	**0.10**/**0.50**	**0.40**/**0.38**	0/0	-0.25/0	0/0	0/0	0/0
	ANCA	**0.86**/**0.80**	**0.63**/**0.88**	0/0	-0.50/-0.88	0/0	0/0	-0.80/-0.44
	MPGN	-0.50/-0.73	-0.29/-0.50	-0.51/-0.51	0/-0.50	**0.17**/0	-0.50/-0.89	-0.50/-0.60
	AL	0.22/0.27	**0.24**/**0.16**	0/0	0/0	0/0	0/0	0/0

3 Proposed Method

In this section, we introduce a customized knowledge distillation method for nephropathy diagnosis over IF sequence. Figure 3 illustrates the overall framework of the proposed method. As shown in the figure, a student network $\mathrm{N_s}$ is constructed to diagnose the nephropathy from the input IF sequence with multiple modalities, which is same as the piratical scenario. Besides, we further develop M teacher networks $\{\mathrm{N_t^i}\}_{i=1}^{M}$ with the similar structure as the student network, but were trained over each single IF modality. During the training on the student network, given a input IF sequence, a learnable customized recruitment module is developed to adaptively select and fuse the knowledge (i.e., the intermediate features and diagnosis logits) from teacher networks, based on the

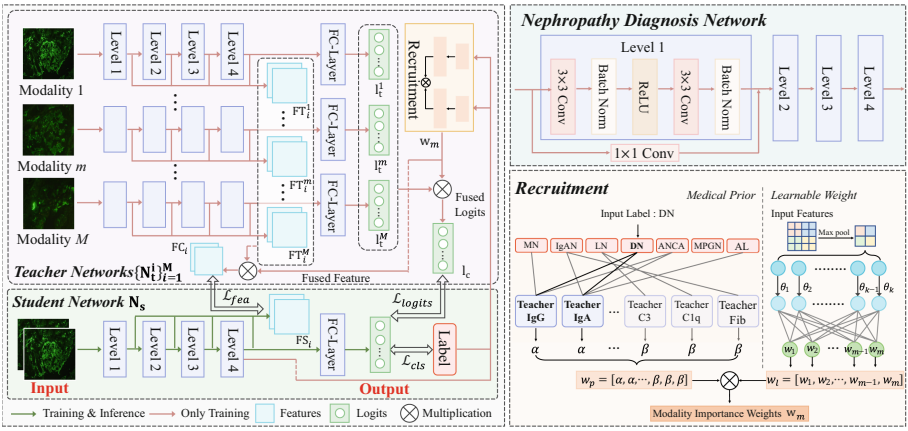

Fig. 3. The overall architecture of the proposed framework.

medical prior from our findings. Then, we transfer the fused knowledge to optimize the student network, via the developed multi-level distillation losses. This way, the student network can dynamically learn from the individual modalities, and finally achieve much better performance overs multi-modality IF sequence.

3.1 Nephropathy Diagnosis Network

Here, we introduce the detailed structure of the nephropathy diagnosis network, as the backbone structure for both student and teacher networks. Note that the backbone structure is flexible, and we implement 4 advanced DNNs in the experimental section. Taking ResNet-18 for the student network as an example, as illustrated in Fig. 3, ResNet-18 is implemented by 4 residual blocks, to extract the features with multiple levels. Each residual block consists of two 3×3 convolutional layers, two batch normalization layers and a ReLU activation layer. Given the input IF sequence $\mathbf{X}$, the multi-level features $\{\mathbf{FS}_i\}_{i=1}^4$ and prediction logits $\mathbf{l_s}$ of the student network $\mathrm{N_s}(\cdot)$ can be obtained as

$$\{\mathbf{FS}_i\}_{i=1}^4, \mathbf{l_s} = \mathrm{N_s}(\mathbf{X}), \mathbf{FS}_{i+1} = \mathrm{Res}(\mathbf{FS}_i) + \mathrm{Conv}_{1\times1}(\mathbf{FS}_i), \mathbf{l_s} = \mathrm{MLP}(\mathbf{FS}_4). \quad (2)$$

In (7), $\mathrm{Res}(\cdot)$, $\mathrm{Conv}_{1\times1}(\cdot)$, and $\mathrm{MLP}(\cdot)$ indicate the residual block, 1×1 convolutional layer, and multilayer perceptron, respectively. As the supervision of N-class nephropathy diagnosis, the cross-entropy loss is calculated upon the one-hot ground-truth diagnosis label $\hat{\mathbf{l}}_\mathrm{s}$ and the output predicted logits $\mathbf{l_s}$:

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{n=1}^{N} \hat{l}_s^n \log l_s^n. \quad (3)$$

3.2 Customized Recruitment Module

A customized recruitment module is developed to adaptively select the effective teacher networks, on the top of the medical priors from our findings. As shown

in Fig. 3, the recruitment module is composed with medical and learnable parts. For the medical prior part, we first construct the adjacency matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, in which M and N indicate the number of IF modalities and nephropathy. In $\mathbf{A}$, the corresponding element is set as 1, when the IF modality is found to have positive influence over 2 DNN models in Table 1. Then, given the ground-truth nephropathy label $\hat{\mathbf{l}}_{\mathrm{s}} \in \mathbb{R}^{N \times 1}$, the medical prior weights can be obtained as:

$$\mathbf{w_p} = \alpha(\mathbf{A} \cdot \hat{\mathbf{l}}_{\mathrm{s}}) + \beta, \tag{4}$$

where α and β are rescaling hyper-parameters. Additional, for the learnable part, the last level feature of the student network $\mathbf{FS}_4$ is passed through a max pooling layer to obtain the representation $\mathbf{v}_{\mathrm{s}} \in \mathbb{R}^{K \times 1}$ for student network, in which K is channel number of $\mathbf{FS}_4$. Let $\boldsymbol{\theta}$ and $\{\mathbf{v}_{\mathrm{t,i}}\}_{i=1}^{M}$ denote the learnable k-element rescaling weights and M teacher network representations. Then, the correlation between student network and teacher network can be formulated as the inner product between vector representations $\langle \mathbf{v}_{\mathrm{t,i}}, \boldsymbol{\theta} \odot \mathbf{v}_{\mathrm{s}} \rangle$. In summary, the learnable importance weights $\mathbf{w_l}$ can be presented as

$$\mathbf{w_l} = \mathbf{V}(\boldsymbol{\theta} \odot (\mathrm{Maxpooling}(\mathbf{FS_4}))), \tag{5}$$

where $\mathbf{V}$ is the matrix of $\{\mathbf{v}_{\mathrm{t,i}}\}_{i=1}^{M}$, while $\odot$ denotes element-wise multiplication. Note that $\mathbf{V}$ and $\boldsymbol{\theta}$ are initialized with 1, and optimized during training. Finally, a overall modality importance weight $\mathbf{w_t}$ can be obtained as

$$\mathbf{w_m} = \mathbf{w_p} \odot \mathbf{w_l}. \tag{6}$$

3.3 Multi-level Knowledge Distillation

Here, a multi-level knowledge distillation is developed between teacher and student networks. Based on the modality importance weight $\mathbf{w_t}$ from our recruitment module, the multi-level features $\{\mathbf{FC}_i\}_{i=1}^{M}$ and predicted logits $\mathbf{l_c}$ of are fused from multiple teacher networks:

$$\mathbf{l_c} = (\sum\nolimits_{j=1}^{M} w_t^j \mathbf{l_t^j})/M, \qquad \mathbf{FC}_i = \sum\nolimits_{j=1}^{M} w_t^j \mathbf{FT}_i^j / M \quad (i = 1, 2, 3, 4). \tag{7}$$

In above equation, w_t^j, $\mathbf{l_t^j}$, $\mathbf{FT}_i^j$ are the importance weight, predicted logits, and the i-th level features for the j-th teacher networks, respectively. After that, to transfer the learned knowledge from teacher to student network, the logit loss is introduced by calculating the Kullback-Leibler (KL) divergence between the fused $\mathbf{l_c}$ and predicted logits $\mathbf{l_s}$ from the student network:

$$\mathcal{L}_{\mathrm{logits}} = \sum\nolimits_{n=1}^{N} \mathbf{l_c^n} \log(\frac{\mathbf{l_c^n}}{\mathbf{l_s^n}}). \tag{8}$$

Meanwhile, in order to make the student network fully learn from diagnosis processes of teacher networks, mean square error (MSE) losses are conducted on each level of fused $\{\mathbf{FC}_i\}_{i=1}^{4}$ and student network features $\{\mathbf{FS}_i\}_{i=1}^{4}$:

$$\mathcal{L}_{\mathrm{fea}} = \sum\nolimits_{i=1}^{4} \|\mathbf{FC}_i - \mathbf{FS}_i\|_2^2. \tag{9}$$

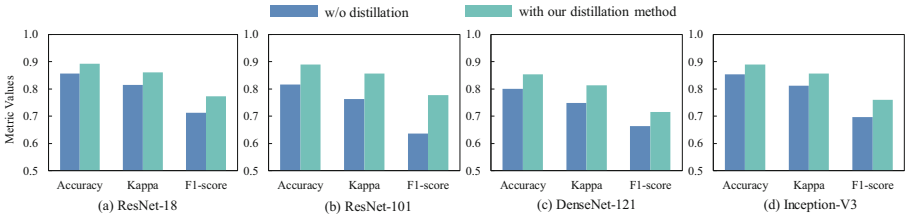

Fig. 4. Diagnosis results of 4 backbone models with our distillation method.

Finally, the overall loss function for the student network can be written as

$$\mathcal{L} = \lambda_{fea}\mathcal{L}_{\text{fea}} + \lambda_{logits}\mathcal{L}_{\text{logits}} + \lambda_{cls}\mathcal{L}_{\text{cls}}, \tag{10}$$

where λ_{fea}, λ_{logits} and λ_{cls} are the hyper-parameters for balancing single losses.

4 Experiment

4.1 Experimental Settings

In our experiments, all IF images are resized to 512×512 for consistency. During training, the parameters are updated by Adam optimizer with an initial learning rate of 0.0001. Then, each teacher network is pre-trained for 70 epochs, and then the student network with our distillation method is trained for 460 epochs. Finally, our and 8 other compared methods are trained over the training set of main set, and evaluated over the main test set and the external set, by adopting 3 evaluation metrics of accuracy, kappa and F1-score.

4.2 Evaluation on Knowledge Distillation

To evaluate the effectiveness of the proposed customized knowledge distillation method, we implement it over 4 different backbone models of ResNet-18 [4], ResNet-101 [4], DenseNet-121 [6] and Inception-V3 [12]. Figure 4 compares the nephropathy diagnosis results of 4 backbone models after conducting our distillation method. As shown, all backbone models obtain significant improvements when applying the proposed distillation methods. For instance, benefiting the transferred knowledge from single-modality, DenseNet-121 improves 0.053, 0.065, and 0.051 in accuracy, kappa and F1-score over our main dataset. This validates the effectiveness of the proposed method. Note that, we select ResNet-18 as the final model, due to its best performance among 4 backbone structures.

4.3 Comparisons with the State-of-the-Art Models

We evaluate the diagnosis performance of our method over our main dataset, compared with 10 DNN based methods, i.e., ShuffleNet [19], EfficientNet [14], GCNet [1], ECANet [15], KNet [7], MANet [18], Hao *et al.* [3], Ada-CCFNet

Table 2. Mean (standard deviation) values of 3 metrics of our and compared methods over our main and external datasets. The best and second best results are in **bold** and underline.

Method	Evaluation on main dataset			Evaluation on external dataset		
	Accuracy	Kappa	F1-score	Accuracy	Kappa	F1-score
ResNet-18 [4]	0.857(0.03)	0.831(0.01)	0.721(0.01)	0.525(0.02)	0.424(0.03)	0.353(0.03)
ResNet-101 [4]	0.813(0.01)	0.760(0.01)	0.620(0.02)	0.673(0.02)	0.600(0.02)	0.586(0.05)
ShuffleNet [19]	0.833(0.01)	0.785(0.01)	0.691(0.02)	0.714(0.04)	0.645(0.05)	0.649(0.04)
EfficientNet [14]	0.841(0.01)	0.795(0.01)	0.676(0.01)	0.437(0.04)	0.306(0.03)	0.264(0.05)
GCNet [1]	0.839(0.01)	0.794(0.01)	0.683(0.02)	0.666(0.03)	0.590(0.03)	0.535(0.03)
DenseNet-121 [6]	0.842(0.01)	0.797(0.01)	0.720(0.01)	0.471(0.04)	0.413(0.03)	0.424(0.03)
ECANet [15]	0.833(0.01)	0.786(0.01)	0.637(0.01)	0.634(0.03)	0.544(0.04)	0.506(0.05)
Inception-V3 [12]	0.855(0.01)	0.817(0.01)	0.709(0.01)	0.709(0.02)	0.646(0.03)	0.605(0.02)
KNet [7]	0.850(0.01)	0.807(0.01)	0.750(0.01)	0.668(0.02)	0.595(0.03)	0.623(0.04)
MANet [18]	0.861(0.01)	0.820(0.01)	0.678(0.02)	0.702(0.03)	0.634(0.04)	0.624(0.02)
Hao *et al.* [3]	0.842(0.01)	0.802(0.01)	0.739(0.02)	0.666(0.03)	0.586(0.04)	0.569(0.02)
Ada-CCFNet [16]	0.851(0.01)	0.820(0.01)	0.669(0.02)	0.678(0.02)	0.521(0.03)	0.520(0.02)
MCL [17]	0.842(0.01)	0.803(0.01)	0.661(0.01)	0.528(0.02)	0.496(0.03)	0.436(0.02)
ITRD [10]	0.806(0.01)	0.773(0.01)	0.646(0.02)	0.618(0.02)	0.593(0.03)	0.507(0.02)
Ours	**0.892(0.01)**	**0.861(0.01)**	**0.773(0.01)**	**0.757(0.02)**	**0.706(0.01)**	**0.655(0.01)**

[16], MCL [17] and ITRD [10]. Among them, KNet [7], MANet [18], Hao *et al.* [3] and Ada-CCFNet [16] are nephropathy diagnosis methods, while MCL [17] and ITRD [10] are knowledge distillation methods. All compared methods are re-trained with the same settings as ours. As shown in Table 2, our proposed method achieves the best performance on IF based nephropathy diagnosis over 3 metrics and 2 datasets. For example, compared with the second best method, our method can improve 0.035/0.043, 0.030/0.060 and 0.023/0.006 in accuracy, kappa and F1-score over our main/external dataset, respectively. This validates the superior performance and generalization ability of the proposed method.

4.4 Ablation Study

We ablate different components of our method to thoroughly analyze their effects on nephropathy diagnosis. Specifically, the accuracy degrades 0.028, 0.023, and 0.031, when ablating medical prior, learnable weights, and recruitment module (equal distillation), respectively. Besides, other ablations, such as the number of teacher networks and distillation loss, are also analyzed. The ablation experiments are reported in Fig. reffinding1 of the supplementary.

5 Conclusion

In this paper, we propose a customized knowledge distillation method for IF based nephropathy diagnosis. Different from the existing methods that averagely integrate information of different IF modalities, we propose a knowledge

distillation framework to transfer knowledge from the trained single-modality teacher networks to a multi-modality student network. In particular, a recruitment module and multi-level knowledge distillation are developed to dynamically select and fuse the knowledge from teacher networks. The extensive experiments on several backbone networks verify the effectiveness of our proposed framework.

References

1. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
2. Gomariz, A., et al.: Modality attention and sampling enables deep learning with heterogeneous marker combinations in fluorescence microscopy. Nature Mach. Intell. **3**(9), 799–811 (2021)
3. Hao, F., Liu, X., Li, M., Han, W.: Accurate kidney pathological image classification method based on deep learning and multi-modal fusion method with application to membranous nephropathy. Life **13**(2), 399 (2023)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2(7) (2015)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Kitamura, S., Takahashi, K., Sang, Y., Fukushima, K., Tsuji, K., Wada, J.: Deep learning could diagnose diabetic nephropathy with renal pathological immunofluorescent images. Diagnostics **10**(7), 466 (2020)
8. Ligabue, G., et al.: Evaluation of the classification accuracy of the kidney biopsy direct immunofluorescence through convolutional neural networks. Clin. J. Am. Soc. Nephrol. **15**(10), 1445–1454 (2020)
9. Liu, Y., Cao, J., Li, B., Hu, W., Maybank, S.: Learning to explore distillability and sparsability: a joint framework for model compression. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
10. Miles, R., Lopez-Rodriguez, A., Mikolajczyk, K.: Information theoretic representation distillation
11. Romagnani, P., et al.: Chronic kidney disease. Nat. Rev. Dis. Primers. **3**(1), 1–24 (2017)
12. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
13. Takahashi, K., Kitamura, S., Fukushima, K., Sang, Y., Tsuji, K., Wada, J.: The resolution of immunofluorescent pathological images affects diagnosis for not only artificial intelligence but also human. J. Nephropathol. **10**(3), e26 (2021)
14. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
15. Wang, Q., Wu, B., Zhu, P.F., Li, P., Zuo, W., Hu, Q.: Eca-net: efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539 (2019)

16. Wang, R., et al.: Ada-ccfnet: classification of multimodal direct immunofluorescence images for membranous nephropathy via adaptive weighted confidence calibration fusion network. Eng. Appl. Artif. Intell. **117**, 105637 (2023)
17. Yang, C., An, Z., Cai, L., Xu, Y.: Mutual contrastive learning for visual representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3045–3053 (2022)
18. Zhang, L., et al.: Classification of renal biopsy direct immunofluorescence image using multiple attention convolutional neural network. Comput. Methods Programs Biomed. **214**, 106532 (2022)
19. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
20. Zhou, X., Laszik, Z., Silva, F.: Algorithmic approach to the interpretation of renal biopsy. In: Zhou, X.J., Laszik Z., Nadasdy, T., D'Agati, V.D., Silva, F.G. (eds.) Silva's Diagnostic Reanal Pathology, pp. pp. 55–57. Cambridge University Press, New York (2009)