Uncertainty-Aware Multi-Expert Knowledge Distillation for Imbalanced Disease Grading

Shuo Tong[1,3,5,*], Shangde Gao[2,3,4,5,*], Ke Liu[2], Zihang Huang[6], Hongxia Xu[4], Haochao Ying[1,3,✉], and Jian Wu[1,3,5]

[1] School of Public Health, Zhejiang University, Hangzhou, China
[2] Computer Sciences and Technology, Zhejiang University, Hangzhou, China
[3] State Key Laboratory of Transvascular Implantation Devices, The Second Affiliated Hospital Zhejiang University School of Medicine, Hangzhou, China
[4] Liangzhu Laboratory and WeDoctor Cloud, Hangzhou, China
[5] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, China
[6] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
{tongshuo, gaosde, lk2017, Einstein, haochaoying, wujian2000}@zju.edu.cn
https://github.com/aTongs1/UMKD

Abstract. Automatic disease image grading is a significant application of artificial intelligence for healthcare, enabling faster and more accurate patient assessments. However, domain shifts, which are exacerbated by data imbalance, introduce bias into the model, posing deployment difficulties in clinical applications. To address the problem, we propose a novel **U**ncertainty-aware **M**ulti-experts **K**nowledge **D**istillation (UMKD) framework to transfer knowledge from multiple expert models to a single student model. Specifically, to extract discriminative features, UMKD decouples task-agnostic and task-specific features with shallow and compact feature alignment in the feature space. At the output space, an uncertainty-aware decoupled distillation (UDD) mechanism dynamically adjusts knowledge transfer weights based on expert model uncertainties, ensuring robust and reliable distillation. Additionally, UMKD also tackles the problems of model architecture heterogeneity and distribution discrepancies between source and target domains, which are inadequately tackled by previous KD approaches. Extensive experiments on histology prostate grading (*SICAPv2*) and fundus image grading (*APTOS*) demonstrate that UMKD achieves a new state-of-the-art in both source-imbalanced and target-imbalanced scenarios, offering a robust and practical solution for real-world disease image grading. The source code has been released by https://github.com/aTongs1/UMKD

Keywords: Knowledge Distillation · Disease Grading · Imbalanced Data

✉: Corresponding author: haochaoying@zju.edu.cn.
*: These authors contributed equally to this work.

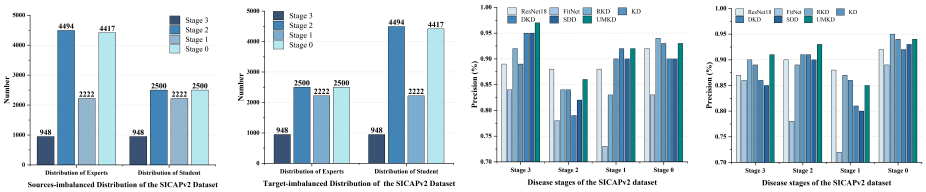

Fig. 1: Domain shifts between source and target data (left) and the performance of methods (right) for *sources-imbalanced* and *target-imbalanced* KD tasks.

1 Introduction

Image-driven disease grading systems are pivotal for enhancing clinical decision-making efficiency [26,14,2], especially for diabetic retinopathy (DR) and prostate cancer. Early-stage precise grading significantly improves patient prognosis (*e.g.*, timely intervention reduces blindness risk by 90% in DR patients) [20,1]. However, traditional grading is largely limited by challenges such as differences in subjective expert judgment and difficulty in identifying subtle pathologic features. Clinical evidence reports a 40% inter-observer variability in Gleason scoring and a misdiagnosis rate of more than 25% for early DR microaneurysms.

Recently, medical efficiency has been enhanced by the rapid development of artificial intelligence-based automatic disease grading systems [3,7,24]. For DR grading, Dai *et al.* [5] developed DeepDR to detect early-to-late stages of diabetic retinopathy. Wang *et al.* [23] reformulate DR grading as sequence prediction, effectively resolving ambiguous boundary issues. On the other hand, for prostate cancer grading, Morpho-Grader [4] disentangles glandular morphology from stromal textures. BayeSeg [9,10] employs variational inference to separate structural invariants from texture variations, significantly improving model robustness. Despite progress in disease grading methods [18,1,15], their deployment in the clinical applications is still limited by domain shifts. Especially, imbalanced data exacerbate domain shifts (differences between source and target domain distributions). As demonstrated in the SICAPv2 dataset (Fig.1 left), stage III prostate cancer samples constitute 8% of the cohort, leading models to overfit to the majority class. This imbalance is equally pronounced in DR grading, where scarce early-stage lesions reduce microaneurysm detection sensitivity by over 30% [16]. Multi-expert knowledge distillation (MKD) [6,11,8], a technique that improves the generalization of the student for minority class samples by transferring expert model knowledge. Due to its robustness to domain shifts, MKD has been applied to address class imbalance in natural images [25,17], but its study in disease image grading remains underexplored.

In this paper, we propose a novel **U**ncertainty-aware **M**ulti-experts **K**nowledge **D**istillation (UMKD) framework to tackle the problem of class imbalance. Specifically, to decouple the structural and semantic information of image representation, we design two feature alignment mechanisms: shallow feature alignment

(SFA) and compact feature alignment (CFA). SFA generalizes the alignment between expert and student features by multi-scale low-pass filtering, thereby preserving structural information (task-agnostic features) of disease images. CFA maps the features of the expert and student models to a common spherical space, allowing the student model to learn grading-related feature knowledge from each expert. We design an uncertainty-aware decoupled distillation (UDD) mechanism at the output space, automatically detecting uncertainties in the expert model caused by class imbalance. Via uncertainty metrics, the student model dynamically adjusts knowledge transfer weights, reducing bias propagation and ensuring a more robust and reliable knowledge transfer process. Experimental results demonstrate that our method significantly outperforms existing multi-expert distillation approaches in both fundus and prostate disease image grading tasks. Particularly for imbalanced classes and heterogeneous models, UMKD achieves more reliable knowledge transfer, as visualized in Fig. 1.

2 Method

Model of the uncertainty-aware multi-expert knowledge distillation. We aim to distill the knowledge from multi-expert models into a target student model for imbalanced disease grading tasks. As shown in Fig. 2, our framework includes shallow feature alignment (SFA), compact feature alignment (CFA), and uncertainty-aware decoupled distillation (UDD). First, we will introduce the feature alignment loss, which will be reused by the SFA and CFA modules. Next, a detailed description of each module will be provided.

Maximum Mean Discrepancy and Reconstruction Loss. To measure the distribution differences between the student's features and those of each expert for feature alignment in the feature space, we employ Maximum Mean Discrepancy (MMD) [6]. The MMD distance is calculated as follows:

$$\mathcal{L}_{\mathrm{MMD}} = \frac{1}{B}\sum_{t=1}^{N}\left\|\sum_{i=1}^{B}\phi\left(\hat{F}_{T_t}^{i}\right)-\sum_{j=1}^{B}\phi\left(\hat{F}_{S}^{j}\right)\right\|_2^2, \tag{1}$$

where ϕ is an explicit mapping function, $\hat{F}_{T_t}$ and $\hat{F}_S$ represent the features of expert T_t and student S after projection, respectively, B is the batch size, and N is the number of experts. Meanwhile, to ensure that the expert models remain unchanged due to privacy constraints, the reconstruction loss $\mathcal{L}_{\mathrm{MSE}}$ is used to measure the changes in the expert models before and after feature alignment:

$$\mathcal{L}_{\mathrm{MSE}} = \sum_{t=1}^{N}\left\|F_{T_t}-\hat{F}_{T_t}\right\|_2^2, \tag{2}$$

where F_{T_t} represents the original features of the expert model T_t before alignment, and $\hat{F}_{T_t}$ denotes the decoded features of the expert model T_t after alignment. By aggregating the alignment loss and the reconstruction loss, the total loss for feature alignment can be expressed as:

$$\mathcal{L}_{\mathrm{FA}} = \mathcal{L}_{\mathrm{MMD}} + \mathcal{L}_{\mathrm{MSE}}. \tag{3}$$

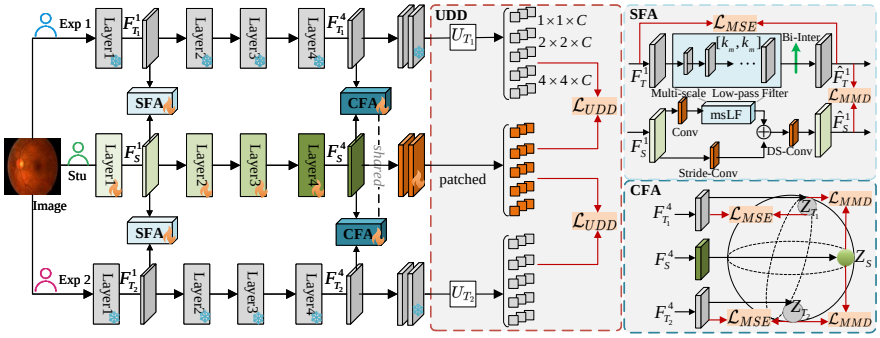

Fig. 2: Model of uncertainty-aware multi-expert knowledge distillation.

Shallow Feature Alignment. SFA preserves the main structural information of given images while removing noise and high-frequency details, ensuring consistency in generalized feature learning between expert models and student models [11]. Specifically, given shallow-layer feature representations between heterogeneous expert-student models, we propose to align features with multi-scale low-pass filtering from a frequency domain. For each expert feature F_{T_t}, we adopt traditional average pooling as the low-pass filter and construct multi-scale filters by adjusting different kernel sizes and strides across multiple groups to accommodate different cutoff frequencies. For the m-th group, the frequency-domain features $\hat{F}_{T_t}$ via multi-scale low-pass filter (msLF) can be expressed as:

$$\hat{F}_{T_t} = \text{msLF}(F_{T_t}) = \Phi(\text{AvgPool}_{k_m \times k_m}(F_{T_t})), \tag{4}$$

where $\text{AvgPool}_{k_m \times k_m}$ denotes the average pooling function with a kernel size of $k_m \times k_m$, and $\Phi(\cdot)$ represents the bilinear interpolation operation.

For the student feature F_S, we design a learnable low-pass filter, which consists of a multi-scale low-pass filter, a convolutional downsampling module, and a depthwise separable convolution (DSConv). The learnable student feature $\hat{F}_S$ in frequency domain is expressed as:

$$\hat{F}_S = \text{Conv}_{3\times3}(\text{Concat}[\text{DownSample}_{s\times s}(F_S), \text{msLF}(F_S)]), \tag{5}$$

where $\text{DownSample}_{s\times s}$, Concat and $\text{Conv}_{3\times3}$ indicates the convolutional downsampling module, feature concatenation operation, and 3×3 convolution operation, respectively. After obtaining the frequency-domain teacher expert and student features with msLF transformation, the total loss of SFA using the losses in Eq. (3) is rewritten as: $\mathcal{L}_{\text{SFA}} = \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{MSE}}$.

Compact Feature Alignment. CFA projects the feature set of the fourth layer (the penultimate layer before the fully connected layer) of all models into a compact high-dimensional spherical space $\mathcal{Z}$. In this space, the student model can learn degrading-related hierarchical knowledge from different pre-trained experts through spatial-domain feature alignment. Considering the heterogeneity among models, the output feature dimensions of their encoders may differ. Before

performing CFA, a 1×1 convolutional kernel is appended to the end of each encoder to adjust the output features of different encoders to the same dimension. In the spherical space $\mathcal{Z}$, the total loss for spatial CFA is computed using the MMD and MSE losses as: $\mathcal{L}_{\text{SFA}} = \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{MSE}}$.

Uncertainty-aware Decoupled Distillation. To address the expert bias for the output prediction, we propose an uncertainty-aware distillation module that dynamically transfers both global and local knowledge from each expert to the student network. Given the logits maps $L_{T_t} \in \mathbb{R}^{C\times H\times W}$ and $L_S \in \mathbb{R}^{C\times H\times W}$ from the expert T_t and student S, we apply spatial partitioning $\mathcal{P}(w, w)$ at multiple scales $w \in W = \{1, 2, 4, ..., w_{\max}\}$. For each partitioned cell $Z(w, n)$ at scale w (where $n \in N_w = \{1, 4, 16, ..., w^2\}$), the accumulated logits are:

$$\psi_{T_t}(w,n) = \frac{1}{w^2} \sum_{(j,k)\in Z(w,n)} L_{T_t}(j,k), \quad \psi_S(w,n) = \frac{1}{w^2} \sum_{(j,k)\in Z(w,n)} L_S(j,k). \tag{6}$$

Then, we devise an uncertainty coefficient U_{T_t} that incorporates the teacher's prediction confidence. For each scale-region pair (w,n), building upon the decoupled knowledge distillation paradigm [27], the UDD loss can be defined as:

$$\mathcal{L}_{\text{UDD}}(w,n) = (2 + U_{T_t}) \cdot \mathcal{L}_{\text{TCKD}} + (1 - U_{T_t}) \cdot \mathcal{L}_{\text{NCKD}}. \tag{7}$$

Here, the loss components are defined as: $\mathcal{L}_{\text{TCKD}} = \|\sigma(\psi_{T_t}(w,n)) - \sigma(\psi_S(w,n))\|_2^2$, $\mathcal{L}_{\text{NCKD}} = \|\psi_{T_t}(w,n) - \psi_S(w,n)\|_2^2$. σ denotes the $softmax$ function. $U_{T_t} = 1 - \max(\sigma(\psi_{T_t}(w,n))) \in [0,1]$, which quantifies prediction ambiguity by measuring the deviation from a one-hot distribution. The $(2 + U_{T_t})$ term amplifies supervision on ambiguous regions ($U_{T_t} \to 1$) where expert predictions tend to be unreliable, while $(1 - U_{T_t})$ maintains precise logit alignment for task-agnostic regions ($U_{T_t} \to 0$).

Training Loss Function. The total training loss can be described as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{cls}} + \alpha \cdot (\mathcal{L}_{\text{SFA}} + \mathcal{L}_{\text{CFA}}) + \beta \cdot \sum_{w\in W} \sum_{n\in N_w} \mathcal{L}_{\text{UDD}}(w,n), \tag{8}$$

where $\mathcal{L}_{\text{cls}}$ is the *cross-entropy* loss and α, β balance the loss components.

3 Experiments

3.1 Setups

Datasets and Metrics. We evaluate our UMKD on two widely used datasets: SICAPv2 [22] for histology prostate grading and APTOS [16] for fundus image grading. Four well-organized metrics are adopted for performance comparison: overall accuracy (OA), mean accuracy (mAcc), weighted F1-score (F1), and mean absolute error (MAE). The **bold** and underline indicate the best and the second-best performance in each sub-dataset test.

Implementation details. We conduct all experiments on two real-world challenging tasks: *sources-imbalanced* distillation and *target-imbalanced* distillation. In *sources-imbalanced* distillation, the expert models are trained on imbalanced source datasets, and distillation is performed on class-balanced target

Table 1: Prostate cancer grading using individually trained ResNet models (Top), feature-based KD models (Middle), and KD models (Bottom).

Methods	Source-imbalanced KD (%)				Target-imbalanced KD (%)			
	OA↑	mAcc↑	F1 ↑	MAE ↓	OA↑	mAcc↑	F1 ↑	MAE↓
Resnet50 (Exp_1)	91.53	89.48	91.47	0.1098	89.19	89.44	89.13	0.1332
Resnet50 (Exp_2)	92.05	89.88	91.93	0.1100	89.71	89.78	89.61	0.1322
Resnet18 (Stu)	89.58	89.06	89.49	0.1463	90.36	88.11	90.23	0.1318
FitNet [21]	78.78	78.41	78.42	0.3136	83.01	78.21	82.60	0.2648
RKD [19]	88.54	88.24	88.44	0.1690	90.79	88.30	90.67	0.1369
KD [13]	89.06	88.39	88.97	0.1624	90.97	89.44	90.90	0.1368
DKD [27]	86.91	85.01	86.68	0.1739	89.19	87.11	89.12	0.1476
SDD [25]	87.82	86.66	87.67	0.1594	89.93	88.81	89.85	0.1447
UMKD	**91.02**	**90.23**	**90.94**	**0.1294**	**91.75**	**90.72**	**91.72**	**0.1199**
△	+3.20	+3.57	+3.27	+0.0300	+1.82	+1.91	+1.87	+0.0248

datasets. In *target-imbalanced* distillation, the expert models are trained on balanced source datasets, and distillation is performed on class-imbalanced target datasets. Specifically, we generate balanced subsets from original imbalanced datasets through random sampling, including SICAPv2-balanced (2500, 2222, 2500, 948) and APTOS-balanced (600, 370, 300, 193, 295) for each grading category. The training, validation, and test sets for all datasets are split in an 8:1:1 ratio, and data augmentation techniques such as random cropping and flipping are employed to expand the dataset. Notably, color jittering is excluded due to the sensitivity of pathological images to color variations, as random color injection could disrupt pathological features. For model training, we utilize two ImageNet pre-trained ResNet50 models as expert models and one ImageNet pre-trained ResNet18 model as the student model.

Baselines. We compare UMKD with a variety of the most representative SOTA methods, including feature-based (FitNet [21] and RKD [19]), as well as logits-based (KD [13], DKD [27] and the latest SDD [17,25]). In addition, we also report the results of ResNet [12] trained on each dataset as a benchmark.

3.2 Comparison Results

Results on SICAPv2 Grading. Our UMKD outperforms all previous logits-based and feature-based KD baselines across all metrics, achieving a new state-of-the-art (SOTA) performance as shown in Table 1.

In the *sources-imbalanced* KD task, UMKD achieves the highest overall accuracy (OA = 91.02%), mean accuracy (mAcc = 90.23%), and weighted F1 score (F1 = 90.94%), while also attaining the lowest mean absolute error (MAE = 0.1294). Specifically, compared to SDD [25], UMKD improves the mean accuracy by 3.57%, with all performance gains highlighted in green. Similarly, in the *target-imbalanced* KD task, UMKD achieves SOTA performance, with mean

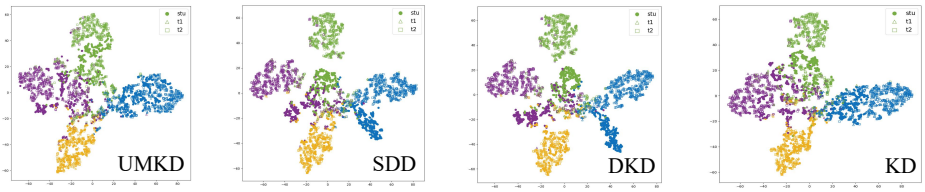

Fig. 3: T-SNE Visualization of different methods on SICAPv2 Grading.

Table 2: Fundus image grading using individually trained ResNet models (Top), feature-based KD models (Middle), and KD models (Bottom).

Methods	Sources-imbalanced KD (%)				Target-imbalanced KD (%)			
	OA↑	mAcc↑	F1 ↑	MAE ↓	OA↑	mAcc↑	F1 ↑	MAE↓
Resnet50 (Exp_1)	82.34	66.33	80.63	0.2478	72.66	63.74	72.18	0.4001
Resnet50 (Exp_2)	82.81	65.77	81.89	0.2421	72.66	63.74	72.51	0.3936
Resnet18 (Sup)	74.21	67.18	73.94	0.4192	82.34	70.56	81.04	0.2521
FitNet [21]	67.57	59.12	66.57	0.5704	79.06	55.77	77.12	0.3694
RKD [19]	74.61	67.15	74.37	0.4203	**85.00**	69.75	**84.38**	0.2482
KD [13]	73.04	66.07	72.77	0.4448	83.43	71.56	83.50	0.2578
DKD [27]	70.70	61.74	69.83	0.4018	81.87	73.95	82.43	0.2743
SDD [25]	73.44	65.07	72.62	0.4017	83.12	73.55	82.83	0.2662
UMKD	**74.61**	**67.33**	**74.43**	**0.3589**	83.91	**74.38**	84.03	**0.2476**
△	+1.17	+2.26	+1.81	+0.0428	+0.79	+0.83	+1.20	+0.0186

accuracy exceeding 90.72%. This consistent superiority of UMKD across both tasks highlights its robustness and generalizability in diverse distillation settings.

Notably, the *sources-imbalanced* KD is more challenging than the *target-imbalanced* task due to the inherent biases in the expert models' knowledge, which are trained on imbalanced datasets. Our UMKD explicitly quantifies and mitigates this imbalance bias, ensuring a more robust and reliable distillation process. In summary, our UMKD not only reduces the number of model parameters (ResNet18 vs. ResNet50) but also significantly enhances model performance, making it a promising tool for improving the accuracy and reliability of prostate cancer grading and diagnosis. To further validate the effectiveness of our method, we visualize the t-SNE results of different approaches. As shown in Fig. 3, our UMKD demonstrates its effectiveness through strong intra-class cohesion and clear inter-class separation, which is consistent with the quantitative results.

Results on APTOS Grading. We evaluate the performance of UMKD on the more challenging APTOS dataset, where data is more imbalanced and expert annotation is biased due to the inherent complexity of fundus images.

As shown in Table 2, UMKD achieves superior performance in both *sources-imbalanced* and *target-imbalanced* KD tasks compared to existing methods. Specif-

Table 3: Ablation study of SFA, CFA, and UDD modules on SICAPv2 dataset.

Methods			Sources-imbalanced KD (%)				Target-imbalanced KD (%)			
SFA	CFA	UDD	OA↑	mAcc↑	F1 ↑	MAE ↓	OA↑	mAcc↑	F1 ↑	MAE↓
			87.82	86.66	87.67	0.1594	89.93	88.81	89.85	0.1447
	✓	✓	90.69	89.92	90.58	0.1314	91.62	90.54	91.58	0.1258
✓		✓	90.36	89.68	90.25	0.1355	91.44	90.93	91.40	0.1275
✓	✓		88.15	86.84	87.97	0.1566	90.06	89.15	90.03	0.1416
✓	✓	✓	**91.02**	**90.23**	**90.94**	**0.1294**	**91.75**	**90.72**	**91.72**	**0.1199**

ically, for *sources-imbalanced* KD task, our proposed UMKD consistently outperforms all baselines, achieving the highest overall accuracy (OA = 74.61%), mean accuracy (mAcc = 67.33%), and weighted F1 score (F1 = 74.43%), while attaining the lowest mean absolute error (MAE = 0.3589). These results demonstrate UMKD's capability to quantify the prediction bias in expert models induced by imbalanced training data. By leveraging the uncertainty of the expert models' outputs, the student model can adaptively adjust the weights of knowledge transfer, thereby ensuring a more robust and reliable distillation process while significantly enhancing model performance.

In the *target-imbalanced* distillation task, UMKD achieves SOTA performance compared to all logits-based methods, attaining the highest mean accuracy (mAcc = 74.38%) and maintaining strong performance across other metrics (OA = 83.91%, F1 = 84.03%). As shown in 7-th row, UMKD yields suboptimal results in OA and F1 compared to the feature-based RKD [19], yet it significantly outperforms RKD in mACC (74.38% vs. 69.75%). We attribute this to the following reasons: first, RKD optimizes the angular momentum to compute mini-batch distances between sample triplets, encouraging samples of the same class to cluster closer together. While this approach achieves the highest overall accuracy, it tends to favor majority classes in *target-imbalanced* distillation tasks due to the inherent class imbalance within batches. In contrast, UMKD addresses this bias by explicitly measuring and mitigating the imbalance with uncertainty, achieving a better trade-off between overall accuracy and mean accuracy.

3.3 Ablation Study

We ablate the contributions of UMKD's three components. Results on the SICAPv2 dataset (Table 3) show that all components are critical for high performance in both *source-imbalanced* and *target-imbalanced* grading tasks. Removing SFA and CFA leads to significant degradation, as task-agnostic structural features and task-specific semantic features cannot be effectively decoupled. This is particularly problematic in disease image grading, where localized pathological features may be obscured by reliance on global features alone. Without UDD, the student model fails to dynamically adjust knowledge transfer weights, resulting in poor distillation performance. These findings underscore the importance

of uncertainty-aware mechanisms for mitigating bias propagation and ensuring robust knowledge transfer. Ablation experiments on the APTOS dataset yield consistent conclusions but are omitted due to space constraints.

4 Conclusion

We proposed a UMKD framework to address the challenge of data imbalance in grading tasks. By integrating the frequency-domain SFA and spatial-domain CFA modules, UMKD effectively decoupled task-agnostic structural features from task-specific semantic features. The UDD mechanism further enhanced robustness by dynamically adjusting knowledge transfer weights based on expert knowledge uncertainties and mitigated biases induced by imbalanced data and model heterogeneity. Extensive experiments on fundus and prostate cancer datasets have demonstrated that UMKD achieved state-of-the-art performance in both source-imbalanced and target-imbalanced scenarios.

Acknowledgments. This research was partially supported by National Natural Science Foundation of China under Grant No. 62476246 and No. 92259202, "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C02132, and GuangZhou City's Key R&D Program of China under Grant No. 2024B01J1301.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)
2. Chen, K., Liu, S., Zhu, T., Qiao, J., Su, Y., et al.: Improving expressivity of gnns with subgraph-specific factor embedded normalization. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 237–249 (2023)
3. Cheng, Y., Ying, H., Hu, R., Wang, J., Zheng, W., Zhang, X., Chen, D., Wu, J.: Robust image ordinal regression with controllable image generation. arXiv preprint arXiv:2305.04213 (2023)
4. Claudio Quiros, A., Coudray, N., Yeaton, A., Yang, X., Liu, B., Le, H., Chiriboga, L., Karimkhan, A., Narula, N., Moore, D.A., et al.: Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. Nature Communications **15**(1), 4596 (2024)
5. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., et al.: A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature communications **12**(1), 3242 (2021)
6. Gao, S., Fu, Y., Liu, K., Gao, W., Xu, H., Wu, J., Han, Y.: Collaborative knowledge amalgamation: Preserving discriminability and transferability in unsupervised learning. Information Sciences **669**, 120564 (2024)

7. Gao, S., Fu, Y., Liu, K., Han, Y.: Contrastive knowledge amalgamation for unsupervised image classification. In: International Conference on Artificial Neural Networks. pp. 192–204. Springer (2023)
8. Gao, S., Fu, Y., Liu, K., Xu, H., Wu, J.: Ka 2 er: Knowledge adaptive amalgamation of experts for medical images segmentation. In: MICCAI Challenge on Comprehensive Analysis and Computing of Real-World Medical Images, pp. 202–214. Springer (2024)
9. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
10. Gao, S., Zhuang, X.: Bayesian image super-resolution with deep modeling of image statistics. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1405–1423 (2022)
11. Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., Xu, C.: One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. Advances in Neural Information Processing Systems **36** (2024)
12. 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)
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
14. Huang, Z., Wang, Z., Zhao, T., Ding, X., Yang, X.: Toward high-quality pseudo masks from noisy or weak annotations for robust medical image segmentation. Neural Networks **181**, 106850 (2025)
15. Huang, Z., Yang, Y., Zhao, T., Yang, X.: A noise robust framework via uncertainty guidance for medical image segmentation with noisy label. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
16. Karthik, Maggie, Dane, S.: Aptos 2019 blindness detection. `https://kaggle.com/competitions/aptos2019-blindness-detection` (2019), kaggle
17. Li, L., Li, X.C., Ye, H.J., Zhan, D.C.: Enhancing class-imbalanced learning with pre-trained guidance through class-conditional knowledge distillation. In: Forty-first International Conference on Machine Learning (2024)
18. Mohan, N.J., Murugan, R., Goel, T., Roy, P.: Drfl: federated learning in diabetic retinopathy grading using fundus images. IEEE Transactions on Parallel and Distributed Systems **34**(6), 1789–1801 (2023)
19. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
20. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. Medical image analysis **59**, 101561 (2020)
21. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
22. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer methods and programs in biomedicine **195**, 105637 (2020)
23. Wang, J., Cheng, Y., Chen, J., Chen, T., Chen, D., Wu, J.: Ord2seq: Regarding ordinal regression as label sequence prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5875 (2023)

24. Wang, J., Xu, Z., Zheng, W., Ying, H., Chen, T., Liu, Z., Chen, D.Z., Yao, K., Wu, J.: A transformer-based knowledge distillation network for cortical cataract grading. IEEE Transactions on Medical Imaging **43**(3), 1089–1101 (2023)
25. Wei, S., Luo, C., Luo, Y.: Scaled decoupled distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15975–15983 (2024)
26. Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S.: A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis **69**, 101985 (2021)
27. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)