scientific reports

Check for updates

OPEN

High efficiency classification of thyroid cytopathological images based on knowledge distillation and vision transformer

Jiazhe Zhang[1], Haolin Zhang[1], Peng Jiang[2], Qin Huang[3], Guangya Zhu[1], Jingjing Chen[4], Yingling Cheng[4], Shu Ran[5] & Fusong Jiang[6]✉

Thyroid cancer is one of the most common types of cancer, pathological diagnosis based on Fine Needle Aspiration Cytology is clinically used as the standard for assessing thyroid cancer. However, the complex structure and large-scale data volume of thyroid pathology images pose challenges in terms of accuracy and efficiency for automatic diagnosis. To address this practical problem, this paper proposes a knowledge distillation method called Multi-Dimensional Knowledge Distillation, which involves feature-based distillation and response-based distillation.We employ a 12-layer Vision Transformer as the teacher model. Feature-based distillation integrates feature information from spatial, channel, and class token, while response-based distillation is achieved through alignment with targets. We integrate information from these diverse dimensions and compress the knowledge into a 3-layer Vision Transformer, which serves as the student model. The student model is trained and evaluated using a dataset containing 22,111 thyroid cytopathological patches. Ultimately, our student model attains a Top-1 classification accuracy of 94.87%. Compared with the teacher model, there is only a 0.55% gap in accuracy, while the computational complexity of the model has decreased by approximately a factor of four. In addition, our method is capable of substantially inheriting the generalization advantages of the teacher model. These results collectively demonstrate the effectiveness of Multi-Dimensional Knowledge Distillation in knowledge transfer.

Keywords Thyroid cancer, Pathological diagnosis, Deep learning, Knowledge distillation

Thyroid nodules are a common disorder that refers to localized lesions in the thyroid gland caused by abnormal, focal growth of thyroid cells. Studies have shown that 7–15% of thyroid nodules may develop into thyroid cancer[1]. According to the National Cancer Report 2022 released by the National Cancer Center, the incidence of thyroid cancer in China is on the rise, especially among the female population[2]. Therefore, how to accurately screen thyroid cancer patients from the thyroid nodule population has become an important challenge in the clinical diagnosis and treatment of thyroid nodules. At present, research endeavors within the realm of thyroid cancer are predominantly focused on modalities such as ultrasonography and molecular testing. However, according to guideline recommendations[3], pathological diagnosis through Fine Needle Aspiration Cytology (FNAC) is the most significant basis for diagnosing thyroid cancer. Specifically, FNAC refers to the ultrasound-guided aspiration of a small number of thyroid cells via fine-needle aspiration for pathological diagnosis. Nevertheless, due to the acute shortage of qualified pathologists and the excessive time consumption involved, the pathological diagnosis of thyroid nodules is currently facing a challenging predicament.

The advancement of deep learning has offered viable approaches for resolving this issue. Researchers can generate Whole Slide Images (WSI) by scanning thyroid cytopathological smears. These WSIs are subsequently employed for the training of deep learning models, enabling intelligent diagnosis. Due to the complex features

[1]College of Engineering Science and Technology, Shanghai Ocean University, Shanghai 201306, China. [2]Department of Information, Shanghai Jiao Tong University School of Medicine Affiliated Sixth People's Hospital, Shanghai 200233, China. [3]Department of Pathology, Shanghai Jiao Tong University School of Medicine Affiliated Sixth People's Hospital, Shanghai 200233, China. [4]Graduate School of Jiangxi, University of Traditional Chinese Medicine, Nanchang 330004, Jiangxi Province, China. [5]School of Health, University of Shanghai for Science and Technology, Shanghai 200093, China. [6]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University School of Medicine Affiliated Sixth People's Hospital, Shanghai 200233, China. ✉email: hajfs@126.com

of thyroid nodule cell pathological images, such as diverse cell morphology, labeling costs, many studies[5–7] limit the identification of thyroid nodules to the binary classification range, namely malignant and benign. In fact, although binary classification models can achieve favorable recognition accuracy, they fail to accurately reflect the degree of malignancy risk for nodules, especially those nodules in the transitional state between benign and malignant, which posing a significant risk of misdiagnosis in clinical practice.To better meet clinical needs, researchers have gradually chosen to use the six risk grades of the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC)[4] as the predictive classes. Wang et al.[8] conducted a large-scale TBSRTC multi-center study on FNAC pathological images of thyroid nodules. Mitsuyoshi et al.[9] established a database comprising 148,395 thyroid cytopathological patches and conducted pathological diagnosis research using EfficientNet. However, their respective researches have consistently revealed that as the predicted categories are further refined, the model's recognition accuracy exhibits a downward trend. Convolutional neural network (CNN) is no longer capable of meeting the classification accuracy demands posed by large-scale image datasets.

In the past few years, deep neural network (DNN) has revolutionized the field of computer vision, such as Vision Transformer (ViT)[10], a landmark achievement of deep neural networks, which has more advantages in terms of accuracy and generalization in the classification of cytopathological images compared to CNN[11–13]. However, the powerful performance of ViT based on the Self-Attention mechanism[10], which leads to high computational and storage costs. This is hard to accept in cytopathology detection that demand both high precision and high efficiency. Therefore, how to lighten the model has become a prerequisite for applying ViT to thyroid pathology diagnosis.

Knowledge Distillation (KD) is a potential method for model compression, aiming to transfer the knowledge of a teacher model into a lightweight student model. The first proposed KD uses the response-based distillation framework[14] to realize knowledge transfer by aligning the output targets of the teacher-student model. As research progresses, researchers have found that relying solely on the target dimension for KD cannot adequately facilitate knowledge transfer between models with substantial differences[15,16]. In addition to the response-based distillation, researchers found that the information extracted from the feature layer of the model can convey more advanced knowledge. For this purpose, FitNets[17] first proposes feature-based distillation, which is subsequently expanded by methods such as Variational Information KD[18], Progressive Blockwise KD[19], Contrastive Representation KD[20]. However, the problems in response-based distillation also occur in feature-based distillation. It is difficult for single-dimensional distillation to meet the lightweight requirements of the teacher-student model with large differences. The relevant research on multi-teacher KD[21] shows that by increasing the number of teacher models and providing the student with learning information from different perspectives, the effect can be improved. The study by Huang et al.[32] found that decoupling the model's targets into inter-class relation and intra-class relation for KD yields superior results. In Relational KD [22], constraining the student model with multiple distance dimensions can also improve the distillation performance. In fact, the limitations of single-dimensional distillation can be explained as the incomplete knowledge transmission of the teacher model and the inability of the student model to fully understand the knowledge.

Based on the issues mentioned above, we propose a method named Multi-Dimensional Knowledge Distillation (Multi-Dimensional KD). This method leverages information from the spatial, channel, class token, and target dimensions to transfer knowledge from a 12-layer ViT to a 3-layer ViT. We utilize a dataset comprising 22,111 thyroid cytopathological patches for model training and validation. The experimental data demonstrate that Multi-Dimensional KD achieves a four-fold reduction in computational complexity at the cost of only a 0.55% decrease in the Top-1 score. Additionally, through visualization analysis, we find that Multi-Dimensional KD can largely preserve the advantages of the teacher model with respect to attention and generalization capabilities. The above results substantiate the effectiveness of our method.

Method

Thyroid nodule cell pathology image dataset

To thoroughly validate the effectiveness of Multi-Dimensional KD, this study collected 600 WSIs of thyroid FNAC specimens from the Department of Endocrinology and Metabolism (Shanghai Sixth People's Hospital) between 2023 and 2024. These WSIs were segmented into 42,032 well-formed patches at 20x magnification using a Leica Aperio AT2 scanner, and were subsequently screened by pathologists. Ultimately, based on the TBSRTC classification, 22,111 patches containing thyroid cytopathologic images from Grade I to Grade VI were labeled, with each patch measuring 224×224 pixels. Of these, 22,111 patches were used for the experiment of the model. Among them, 12,235 patches were employed for the training and testing, while 9,876 patches were utilized for validation.

It is necessary to declare that the collection and processing of all data used in this experiment are in accordance with relevant regulations, and all the parties or their legal guardians have given informed consent. Moreover, the methodological procedures described subsequently are based on previous work experiences[5–7]. Furthermore, we have obtained ethical approval from Shanghai Sixth People's Hospital affiliated to Shanghai Jiao Tong University.

Overall framework

We use ViT as the foundation for KD, where the teacher model comprises 12 transformer blocks, and the student model has 3 transformer blocks. The overview of the proposed Multi-Dimensional KD is shown in Fig. 1, which is divided into two parts: (1) Response-based distillation, which employs a temperature scaling operation to soften hard-targets for calculating the soft-targets loss. The loss between hard-targets and ground truth labels serves as the initial loss for ViT. (2) Feature-based distillation, encompasses spatial-wise KD, which focuses on spatial weight information; Class token KD, which is employed to extract information from the decoupled class

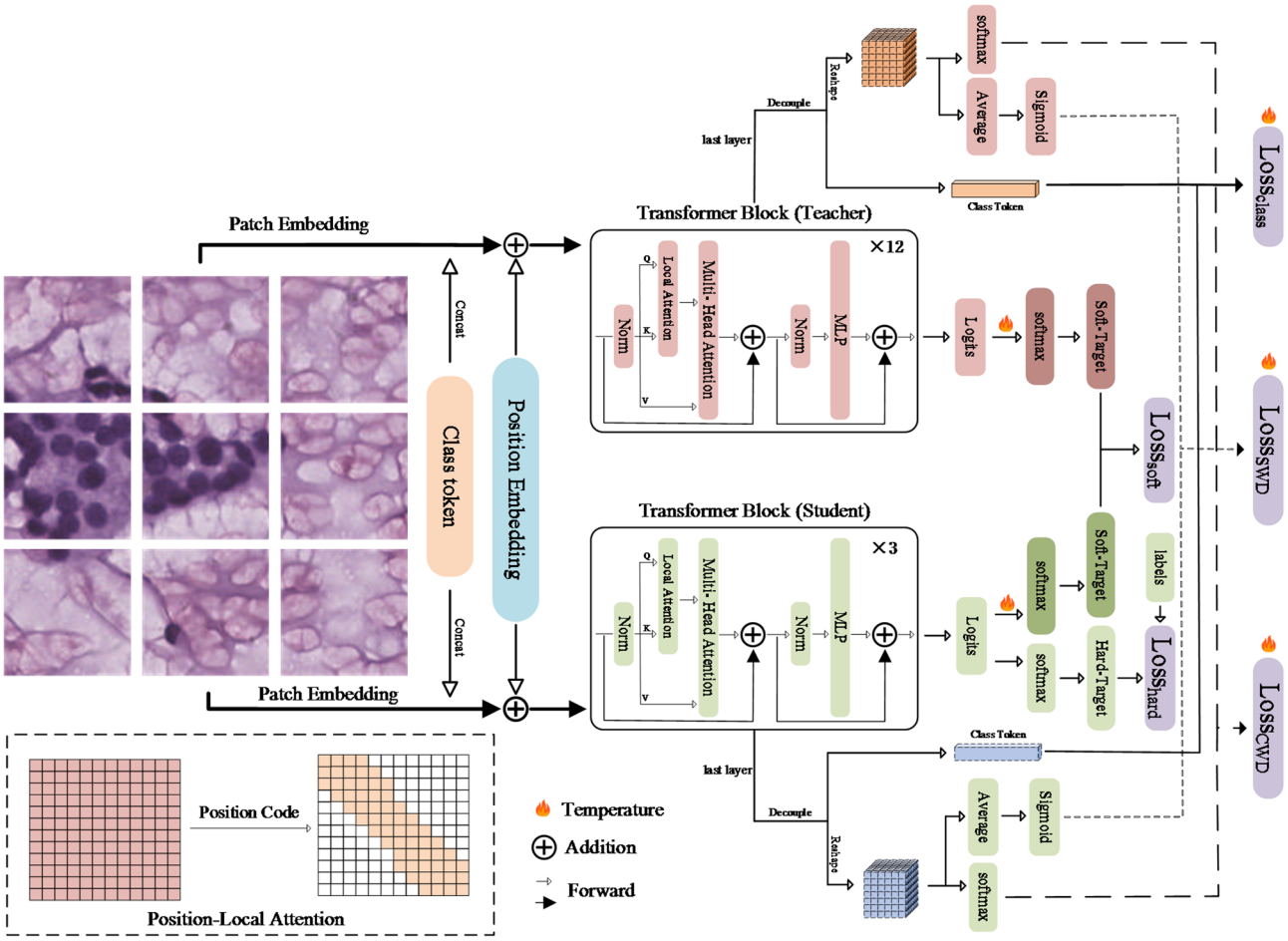

Fig. 1. The overall architecture of our proposed method. Multi-Dimensional KD. Multi-Dimensional KD is applied between two ViT models, where the teacher model consists of 12 transformer blocks, and the student model consists of 3 transformer blocks.

token. Additionally, we employ position-local attention, which incorporates relative position information, to constrain the Self-Attention weight matrix within a banded range.

Response-based distillation

The common denominator in classification problems is that the model output logits. In this paper, logits $\boldsymbol{L^T}$ are obtained from pathological image data through a teacher model that consists of 12 transformer blocks, while logits $\boldsymbol{L^S}$ are derived simultaneously from a student model with only 3 transformer blocks. Typically, when the logits are normalized by the softmax function, a set of hard-targets is generated, each hard-target containing the probabilities for all categories, and the category with the highest probability is selected as the final prediction. But in fact, each probability value in the hard-targets is not zero, and the discarded parts still contain some similarity information. To endow the student model with the ability to extract this information, we replace the hard-targets with a soft-targets using a temperature coefficient $\boldsymbol{t}$, and calculate the Mean Squared Error (MSE) between them to obtain the KD loss at the output layer, which is expressed as:

$$L_{soft} = MSE\left(\theta\ \left(L^S\right), \theta\ \left(L^T\right)\right). \tag{1}$$

Where $\boldsymbol{\theta(\cdot)}$ denotes the softmax function, which is responsible for converting logits into a probability distribution as shown below:

$$\theta\ (L_i) = \frac{e^{\frac{L_i}{t}}}{\sum_{i=1}^{N} e^{\frac{L_i}{t}}} \tag{2}$$

In the formula, $\boldsymbol{i}$ represents the index in the sequence of logits that corresponds to each predicted category, with the maximum value being $\boldsymbol{N}$, which is the number of categories. As the value of $\boldsymbol{t}$ increases, the softmax output exhibits a more homogeneous distribution, which enables the student model to fully extract the hidden

information in targets. Besides the increase in information content, using soft-targets for training also leads to a smaller variance in model gradients, allowing for the use of larger learning rates during training to accelerate convergence. Finally, the soft-targets of the teacher model are taken as the true label, while the soft-targets of the student model are taken as the predicted label. The loss calculation is expressed as:

$$MSE\left(\theta\ \left(L^S\right),\theta\ \left(L^T\right)\right)=\frac{1}{N}\sum{}^{N}_{i=1}(\theta\ \left(L_i^T\right)-\theta\ \left(L_i^S\right))^2 \tag{3}$$

Class token knowledge distillation

The feature map structure of ViT is different from that of CNN, with its uniqueness stemming from class token and Self-Attention weight calculations. Therefore, when performing feature-based distillation on ViT, decoupling the feature maps is first required. The dimensional information is represented as:

$$[Batch\ ,num\_patches+class\_token\ ,embed\_dim]$$

In the second dimension, Class_Token is separated as an embedding representation from the rest, yielding a feature map $\boldsymbol{F}$ with dimensions of (169×768). Subsequently, $\boldsymbol{F}$ is reshaped to (768×14×14) to match the C×W×H format.

In the processing of ViT feature maps, class token[10] learns global information about the entire image within the encoder and is utilized for the final classification task, closely correlating with the model's output targets. Due to the richer information contained in class token, class distillation can be regarded as a dense branch of response-based distillation. Using it as the distillation targets can more evenly distribute the classification loss information across the entire model. We employ Kullback-Leibler divergence (KL divergence) to evaluate the consistency between the two, where $\boldsymbol{\delta}$ represents the vector of the class token:

$$Loss_{class}=\phi\ \left(\theta\ \left(F^S\right),\theta\ \left(F^T\right)\right)=\sum{}^{C}_{c=1}\theta\left(\delta\ {}^{T}_{c}\right)\cdot\ \log\left(\frac{\theta\ \left(\delta\ {}^{T}_{c}\right)}{\theta\left(\delta\ {}^{S}_{c}\right)}\right) \tag{4}$$

Channel-wise knowledge distillation

Each channel of the feature map corresponds to a visual pattern, but the importance of the visual pattern of each channel is different. Since the performance of the teacher is better than that of the student model, we assume that the visual patterns captured by the teacher are more accurate. Therefore, this paper introduced Channel-wise Knowledge Distillation (CWD)[23] for channel-by-channel KD, extracted the attention information of each channel of the feature map with the teacher, and summarized it into knowledge transfer to student models. $\boldsymbol{F^S}$ and $\boldsymbol{F^T}$ are used to respectively represent the feature graphs of the student model and the teacher model, then the loss can be described as:

$$L_{CWD}=\phi\ \left(\theta\ \left(F^S\right),\theta\ \left(F^T\right)\right). \tag{5}$$

Similarly, the softmax function $\boldsymbol{\theta(\cdot)}$ is used to convert feature values into a probability distribution as shown below:

$$\theta\ \left(F_c\right)=\frac{e^{\frac{F_{c,i}}{t}}}{\sum{}^{W\cdot H}_{i=1}e^{\frac{F_{c,i}}{t}}} \tag{6}$$

In this context, $\boldsymbol{c}$ represents the index of each channel in the feature map, and $\boldsymbol{i}$ represents the spatial position within the indexed channel. Similar to the method described in response-based distillation, CWD also adopts temperature-scaling approach. As $\boldsymbol{t}$ increases, the probability distribution becomes smoother, and the differences between positions within a channel decrease, allowing the hidden relationship between channels to be fully revealed. Due to the uniform feature map size in ViT, we use KL divergence to evaluate the difference in channel feature distribution between the teacher model and student model. This is expressed as:

$$\phi\ \left(\theta\ \left(F^S\right),\theta\ \left(F^T\right)\right)=\frac{1}{C}\sum{}^{C}_{c=1}\sum{}^{W\cdot H}_{i=1}\theta\left(F^T_{c,i}\right)\cdot\ \log(\frac{\theta\ \left(F^T_{c,i}\right)}{\theta\left(F^S_{c,i}\right)}) \tag{7}$$

During the distillation process, when the amount of information in the teacher model's feature map is significantly greater than that of the student model, the logarithm function in the KL divergence calculation will yield a value greater than zero, which manifests as a relatively large loss value. As the training progresses, the feature differences between the teacher and student models decrease. Consequently, the logarithm values move closer to zero, resulting in a gradual decrease in the loss value. This process has the effect of suppressing less significant channels, indicating that the student model's visual attention in the channel dimension is converging towards that of the teacher model.

Spatial-wise knowledge distillation

Although CWD replaces the traditional spatial-focused strategy in feature-based distillation, it inevitably neglects spatial features. Several studies[24–26] have shown that a one-sided focus on channel-level distillation

methods can impair the interaction among spatial information, especially for data with small feature regions such as thyroid pathological images. Multi-scale attention information is more capable of guiding the student model to focus on the correct locations. To make up for the gap in spatial knowledge, this study presents Spatial-wise Knowledge Distillation (SWD), inspired by the spatial attention module. It is described as follows:

$$L_{SWD} = MSE\left(\partial\left(F^S\right), \partial\left(F^T\right)\right). \tag{8}$$

In this context, $\partial(\cdot)$ represents a process that compresses the number of feature map channels C down to 1 through calculating the mean of the feature map. Subsequently, the sigmoid function is applied to convert it into a spatial-wise:

$$\partial\ (F) = \frac{1}{1 + e^{\frac{1}{Ct}\sum_{c=1}^{C} F_c}}. \tag{9}$$

Traditional spatial attention uses both the mean and maximum values to generate a 2-channel attention map. However, as a loss function, SWD cannot utilize convolutional operations to further fuse this information. In our experiments, we found that the loss information derived from the maximum value without convolutional operations significantly hindered gradient descent, leading to rapid degradation of the model. Therefore, SWD only uses the relatively smooth mean-based information to constrain spatial differences, ensuring that the loss values are distributed within a reasonable range. Additionally, within $\partial(\cdot)$, there is also a temperature parameter t that achieves a soft distribution of spatial attention. The larger t is, the more evenly the attention is distributed across each spatial location, creating an adversarial relationship between the teacher model and student model, while enhancing the robustness of the models.

Combining Eq. (1) to (9), the total optimization loss is expressed as:

$$L = \lambda L_{hard} + (1 - \lambda)(L_{soft} + L_{CWD} + L_{SWD} + L_{Class}) \tag{10}$$

Where λ is used to control the proportion of the KD loss in the overall loss, and it is suggested to be set at 0.8 in this paper.

Position-local attention

Self-Attention requires calculating the correlation between any two tokens in the sequence, resulting in a quadratic computational complexity for the Attention weight matrix and incurring significant computational costs. Therefore, we introduce the concept of local attention[27] to restrict the calculation of attention weights for each element in the sequence within a certain range. Formally, local attention can be expressed as only calculating the attention between each element and the neighboring n elements. The mathematical formulation of local attention is expressed as:

$$A_i = q_i[k_{i-\frac{n}{2}}, k_{i-\frac{n-2}{2}}, k_{i-\frac{n-4}{2}}, \ldots, k_{i-\frac{n-2n}{2}}] \tag{11}$$

In this context, A_i represents the attention weight matrix formed by the product of the query value (denoted as q) of the i-th element and the key (denoted as k) values of neighboring n elements. This approach limits the computational load while preserving the ability to connect contextual information. In practice, ViT focuses on global data analysis. However, in thyroid nodule cytopathology images, structures such as papillary follicles and colloidal follicles exhibit unique local features, which are also crucial criteria for pathological diagnosis[28]. In this regard, some variants of ViT, such as TNT[29] and Swin Transformer[30], have incorporated locality. In this paper, this role is undertaken by local attention, which also establishes a global field of view through overlapping local regions. Additionally, due to the loss of positional information caused by the reduction of Self-Attention, we propose a relative positional encoding to highlight the positional relationship between q and k, which can be expressed as:

$$Local_i = [\frac{\frac{3n}{2} - i}{2n} A_{i,1}, \frac{\frac{3n}{2} - i - 1}{2n} A_{i,2}, \frac{\frac{3n}{2} - i - 2}{2n} A_{i,3}, \ldots, \frac{\frac{3n}{2} - i - (n-1)}{2n} A_{i,n}] \tag{12}$$

Relative positional encoding derives relative positional parameters based on the positions of q and k when calculating attention weights. The larger the parameter value, the closer the relationship between the two positions.

Knowledge distillation methods for comparison

To validate the effectiveness of Multi-Dimensional KD, we compared our method with KD methods listed below:

- **DKD**[31]: This method decouples classical KD into target class KD and non-target class KD, and calculates losses separately to minimize the differences in the output layer.
- **DIST**[32]: This method focuses on both Inter-class relations and Intra-class relations in the output layer, conducting KD from two perspectives.
- **TAKD**[16]: This method utilizes a medium-sized teacher assistant model to bridge the large gap between the teacher and student models.

In this research, our objective is to compress the teacher model (12-layer ViT) into a student model (3-layer ViT) through KD, aiming to enable the student model's performance to closely match that of the teacher model. Consequently, we evaluate and compare the performance of the student model under different KD approaches, and take the 3-layer ViT that is directly trained without any KD as the baseline model.

Since the feature layer of ViT differs from traditional neural networks, we selected the feature map after separating the class token as the targets for feature-based distillation. Regarding TAKD, we followed the protocol outlined in the literature[16], where the 12-layer ViT is first distilled into a 7-layer ViT via targets, and subsequently it undergoes a further distillation process to be transformed into a 3-layer ViT. As for DKD, we divide the targets into non-target and target classes, and then calculate the KD loss for each class respectively. In addition, We achieve DIST by calculating the losses of Intra-class and Inter-class relations within the target.

We apply these methods to the same dataset. Model training is conducted on a single NVIDIA RTX 4090 graphics card. The batch size is set to 16, the number of training epochs is set to 600, the learning rate is set to 0.01, and we employ the Stochastic Gradient Descent optimizer.

Evaluation metrics

To assess the performance and efficiency of the Multi-Dimensional KD for thyroid nodule cytopathology image recognition, we employ Top-1, Precision, Recall, and F2 Score to evaluate the classification capabilities of all models. Additionally, we calculate the Floating-point Operations Per Second (FLOPs) and Frames Per Second (FPS) to assess the operational efficiency of the models.

Results

Comparison

The final comparative experimental results are shown in Table 1. Compared with the baseline model that is directly trained without KD, all distillation methods can improve the performance of the student, but Multi-Dimensional KD achieves the best scores. Compared to the 12-layer ViT, our method only exhibits a 0.55% performance difference, while significantly reducing the model's computational complexity from 16.89 GMac to 4.31 GMac. Simultaneously, the model weight drops from 327 MB to 84 MB and the time for processing a single image reduces from 0.014s to 0.0037s. It is noteworthy that the Precision, Recall, and F2 Score of our method are relatively close to each other. This indicates that the model is free from the risks of overfitting or underfitting, demonstrating favorable generalization ability. In the context of explaining KD, this can be understood as a relatively thorough knowledge transfer. Clinically, it implies a lower rate of misdiagnosis.

We have generated a confusion matrix for the model's classification results, as shown in Fig. 2a. It can be observed that classification errors are mainly concentrated between Grade III (Atypia of undetermined significance or follicular lesion of undetermined significance) and Grade IV (Follicular neoplasm or suspicious for follicular neoplasm) cells. In fact, due to the high similarity in cytological features between these two grades, even deep ViT models tend to make numerous classification errors. Clinically, repeated FNAC or molecular testing may be required to further confirm the diagnosis[3].

Under different KD loss weight $\boldsymbol{\lambda}$, the distillation temperature is adjusted to $\boldsymbol{t}$. The experimental results are shown in Fig. 2b. It is found that that increasing $\boldsymbol{t}$ within a certain range can make the output distribution softer, thus improving model performance, but if $\boldsymbol{t}$ continues to increase beyond this range, the performance will decline, which can be explained by the model's excessive attention to some small hidden features. When $\boldsymbol{\lambda = 0.6}$ and $\boldsymbol{t = 2}$, the model achieves the best performance.

We utilized Grad-CAM to provide attention visualization for the student model of each KD method, with the results presented in the form of heatmaps in Fig. 3. The attention of the baseline model and the comparative methods tends to exhibit either excessive sparsity or density, which implies that the models lose key information or incorporate noise during the recognition process. In contrast, Multi-Dimensional KD enables the student model to focus its attention on regions densely populated with thyroid cells, demonstrating the model's full utilization of spatial dimensional information. Additionally, the predominantly red colors within the model's high level of attention to channel-dimensional information. All of the above evidence demonstrates our method's precise knowledge transfer capability in both spatial and channel dimensions.

To further validate the performance of Multi-Dimensional KD, we referred to the pseudo-gland test to evaluate the generalization ability of the student model, with the results shown in Fig. 4. For the original images, the student model's attention is well-focused on the regions where thyroid cells aggregated. Subsequently, we add some additional thyroid cells of the same grade to the original images, labeled them as pseudo cells, and had the student model identify them. Through the heatmaps, it can be found that the model's attention could comprehensively cover the pseudo cells, indicating that the model exhibited strong sensitivity to the additional

Method	Top-1	Precision	Recall	F2
Teacher	0.9547	0.9549	0.9558	0.9535
Baseline	0.9168	0.9221	0.9196	0.9183
TAKD	0.9461	0.9475	0.9471	0.9486
DKD	0.9462	0.9469	0.9471	0.9469
DIST	0.9399	0.9430	0.9410	0.9402
Ours	**0.9487**	**0.9501**	**0.9497**	**0.9494**

Table 1. Comparison between Multi-Dimensional KD and other KD methods.

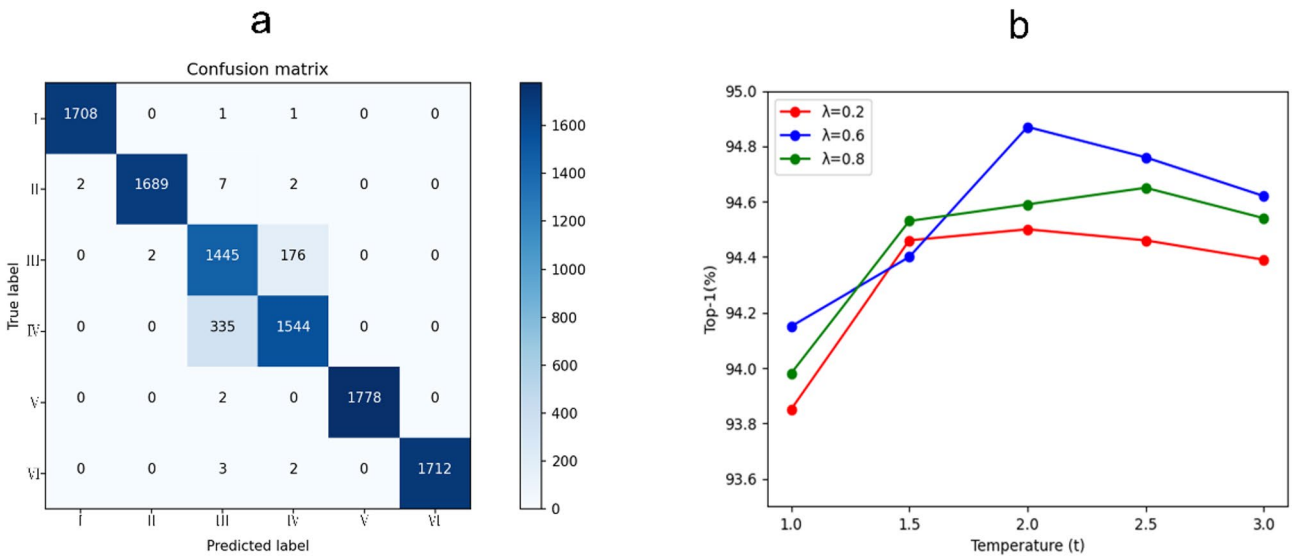

Fig. 2. (**a**) Confusion matrix. (**b**) Top-1 score of the model with different λ and t.

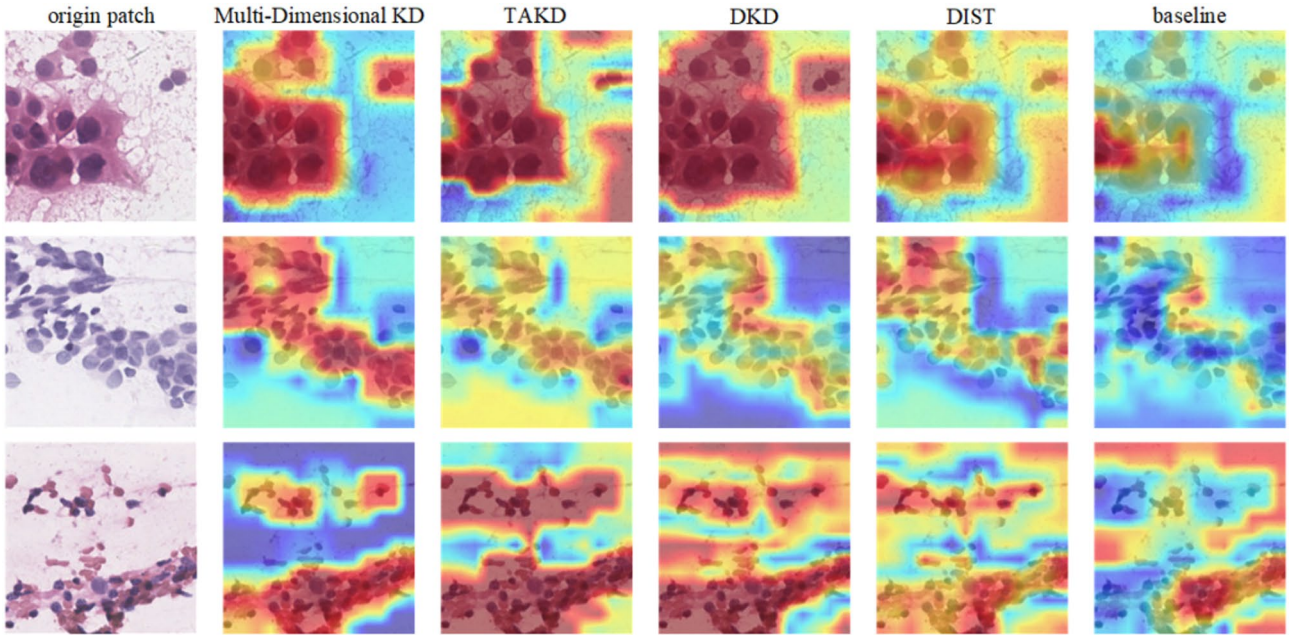

Fig. 3. Visual comparison using Grad-CAM. Grad-CAM shows the contribution of different regions in the input image to the model prediction through a heatmap. The colors in the heatmaps represent the distribution of the model's attention, with a gradient from blue to red indicating a transition from sparse to dense attention distribution.

data. This demonstrates that Multi-Dimensional KD can effectively preserve the generalization advantages of the teacher model.

Ablation study

To validate the contribution of each loss function in enhancing model fitting, we conducted a ablation study. We employed 3-layer ViT as basic student model, equipped with $\boldsymbol{Loss}_{\boldsymbol{hard}}$. Table 2 presents the ablation data for Multi-Dimensional KD. The results indicate that each KD function improves the model's performance, as evidenced by an increase in the Top-1 accuracy from 93.23 to 94.87%, demonstrating the effectiveness of Multi-Dimensional KD in bridging the performance gap between the teacher and student models. Among them, $\boldsymbol{Loss}_{\boldsymbol{CWD}}$ and $\boldsymbol{Loss}_{\boldsymbol{SWD}}$ make the most significant contributions to performance improvement, and this improvement is greater than the sum of their individual contributions. This indicates that there exists both complementary and

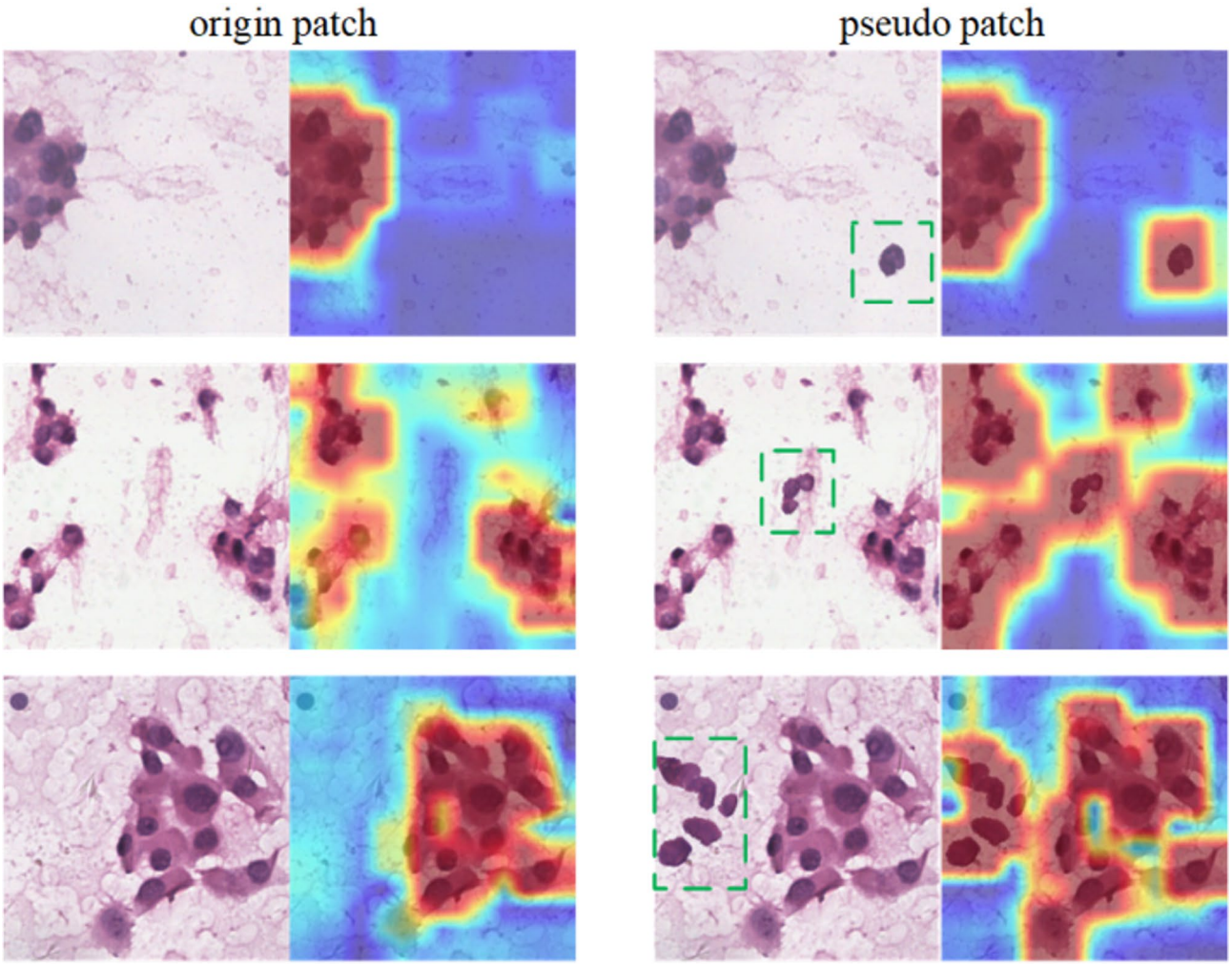

Fig. 4. Generalization Testing. The areas within the green boxes represent the pseudo cells that are manually added to the origin patches.

$Loss_{hard}$	$Loss_{soft}$	$Loss_{CWD}$	$Loss_{SWD}$	$Loss_{class}$	Position-Local Attention	Top−1	FPS
✓						93.23	300
✓	✓					93.94	301
✓	✓	✓				94.55	300
✓	✓		✓			94.20	300
✓	✓			✓		94.51	301
✓	✓	✓	✓	✓		94.95	301
✓	✓	✓	✓	✓	✓	94.87	271

Table 2. Ablation study.

competitive relationships between Channel-wise KD and Spatial-wise KD. When the weights of the two reach a balance, the effectiveness of KD can be maximally enhanced. The incorporation of position-local attention further boosts the model's image processing speed by 10%, with only a slight decrease in Top-1.

We compared the training losses of the basic student model and Multi-Dimensional KD, as illustrated in Fig. 5. We observed that due to the inclusion of more significant differences in feature distributions, the initial loss of Multi-Dimensional KD curve was higher, but the loss of Multi-Dimensional KD decreased at a faster rate during training. This indicating that multi-dimensional supervisory information more effectively promotes the convergence of the student model, and Multi-Dimensional KD has stronger adaptability and generalization to the training data.

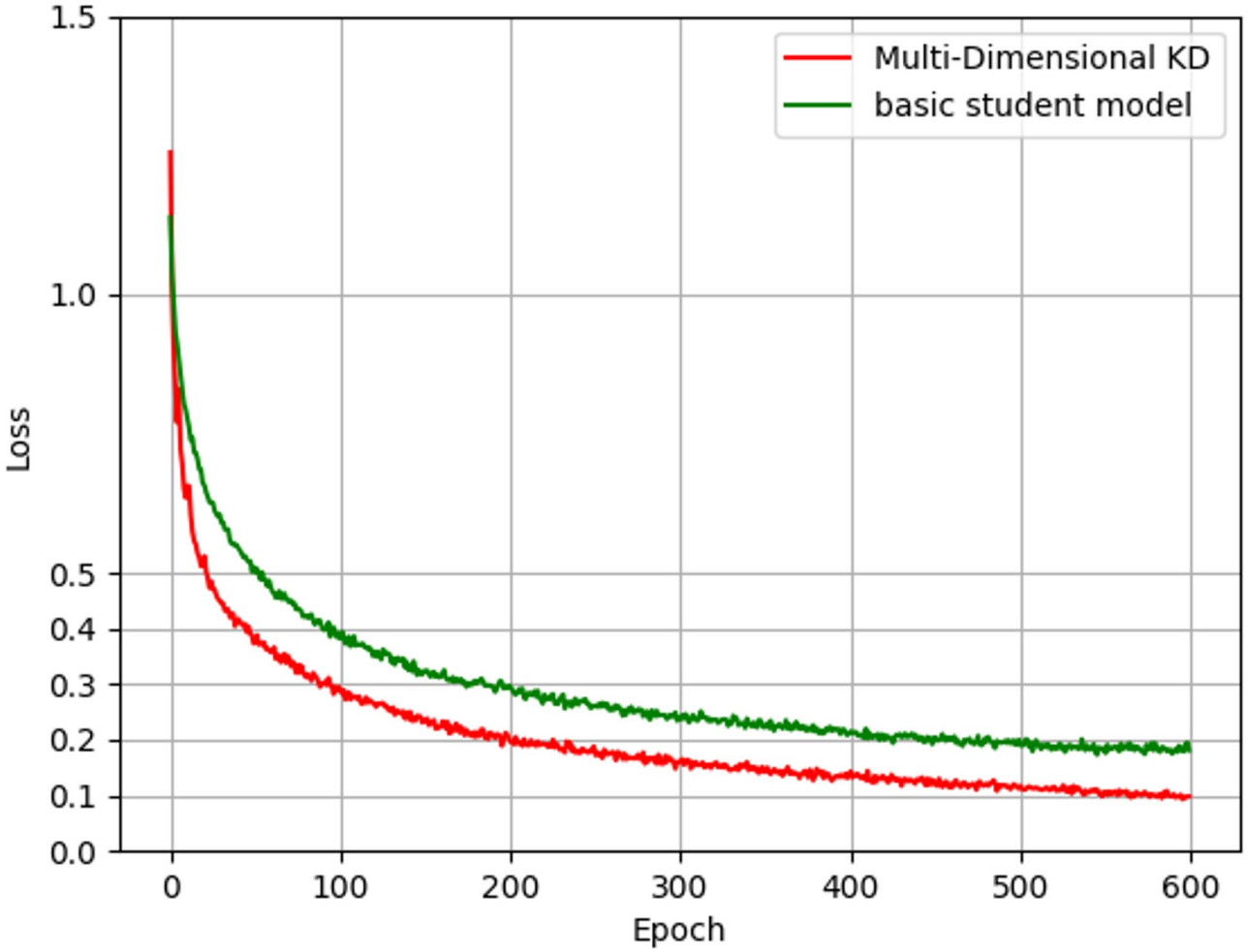

Fig. 5. Training loss curves for the Basic Student Model and Multi-Dimensional KD. Since both student models have the same architecture, we load identical pre-trained weights in both experiments to shorten experimental cycles.

Conclusions

In this study, we propose a novel KD method, Multi-Dimensional KD, which is trained on a dataset comprising 12,235 thyroid cytopathology patches. It facilitates the transfer of knowledge from a 12-layer ViT to a 3-layer ViT, with classification carried out in accordance with TBSRTC. Additionally, we collected an extra 9,876 patches to validate the student model. The experimental results demonstrate that we have successfully reduced the computational complexity of the model from 16.89 GMac to 4.31 GMac, and the recognition time for a single patch has decreased from 0.014 s to 0.0037 s, all while incurring only a minor accuracy loss of 0.55%.

The effectiveness of Multi-Dimensional KD can be attributed to multiple factors. Firstly, during the extraction of target information, the unique architecture of the ViT is leveraged to decouple the class token, which supplements the target and ensures the integrity of the KD information. Secondly, by calculating the channel-wise in the feature maps, Channel-wise KD is achieved, and Spatial-wise KD is employed to compensate for spatial information. As observed from the heatmaps, Multi-Dimensional KD enables the student model's attention to be compactly distributed in the thyroid cell regions and allows it to inherit the generalization advantages of the teacher model. This is precisely due to the synergy between channel-wise and spatial-wise. Thirdly, we utilize a temperature parameter to perform a temperature-raising operation on the entire KD system, making the KD process smoother. This not only reveals hidden information but also enhances the model's robustness. Finally, we introduce local attention to further lighten the model and incorporate relative positional encoding to emphasize positional relationships. As a result, the FPS of the student model decreases by 9.7%.

However, this study has certain limitations. We have found that our method exhibits suboptimal performance in classifying thyroid nodule cytopathology patches of Grade III and Grade IV. In fact, due to their similar morphological characteristics, even in clinical practice, these two grades are distinguished with the assistance of ultrasound and molecular testing. To enhance the potential of our KD system for clinical applications, in the future, we will integrate multi-modal data, including ultrasound and molecular testing results, to address the challenges in pathological recognition. Additionally, although our method can reduce the overall computational load of the model by a factor of four, it still falls short of meeting the requirements for WSI recognition. We have

observed that many regions within a WSI do not contain useful diagnostic information. Therefore, we plan to filter out the ineffective regions of the WSI under low magnification and then apply our method for recognition to fulfill the clinical demands for WSI diagnosis. When dealing with small-scale datasets, we have noticed that the complex architecture of Multi-Dimensional KD is prone to overfitting. To address this issue, we will optimize the loss functions for each dimension at the mathematical level, thereby extending the application scenarios of Multi-Dimensional KD to few-shot detection.

Data availability

Due to the policy of the Sixth People's Hospital of Shanghai Jiao Tong University School of Medicine prohibits public uploads of any patient's private data, the dataset analyzed and generated in this study is not publicly available. However, partial dataset is available from the appropriate authors upon reasonable request. The source code used in this paper has been publicly released at https://github.com/zhezhedepenyou/my-code/tree/main/Multi-Dimensional. Please contact the corresponding author at hajfs@126.com for further information.

Received: 26 February 2025; Accepted: 11 August 2025
Published online: 17 August 2025

References

1. Siegel, R. L. et al. Cancer statistics, 2021. *Cancer J. Clin.* **71** (1), 7–33. https://doi.org/10.3322/caac.21654 (2021).
2. Han, B. et al. Cancer incidence and mortality in china, 2022. *J. Natl. Cancer Cent.* **4** (1), 47–53. https://doi.org/10.1016/j.jncc.2024.01.006 (2024).
3. Tamhane, S. & Gharib, H. Thyroid nodule update on diagnosis and management. *Clin. Diabetes Endocrinol.* **2** (1), 17. https://doi.org/10.1186/s40842-016-0035-7 (2016).
4. Ali, S. Z. et al. The Bethesda system for reporting thyroid cytopathology: definitions, criteria, and explanatory notes. *Springer Int. Publishing* (2023).
5. Dov, D. et al. Deep-Learning-Based screening and ancillary testing for thyroid cytopathology. *Am. J. Pathol.* **193** (9), 1185–1194. https://doi.org/10.1016/j.ajpath.2023.05.011 (2023).
6. Elliott Range, D. D. et al. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. *Cancer Cytopathol.* **128** (4), 287–295. https://doi.org/10.1002/cncy.22238 (2020).
7. Chengwen, D. et al. Differential diagnostic value of the ResNet50, random forest, and DS ensemble models for papillary thyroid carcinoma and other thyroid nodules. *J. Int. Med. Res.* **50** (4), 3000605221094276. https://doi.org/10.1177/03000605221094276 (2022).
8. Wang, J. et al. Deep learning models for thyroid nodules diagnosis of fine-needle aspiration biopsy: a retrospective, prospective, multicentre study in China. *Lancet Digit. Health.* **6** (7), e458–e469. https://doi.org/10.1016/S2589-7500(24)00085-2 (2024).
9. Hirokawa, H. et al. Application of deep learning as an ancillary diagnostic tool for thyroid FNA cytology. *Cancer Cytopathol.* **131** (4), 217–225. https://doi.org/10.1002/cncy.22669 (2023).
10. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at (2020). https://arxiv.org/pdf/2010.11929/1000
11. Xingzhe, C. et al. Method of Risk Stratification for Detecting Malignant C-TIRADS in Thyroid Nodules Based on Self-attention and Self-distillation. Journal of Computer-Aided Design & Computer Graphics. 1–15, (2025). https://doi.org/10.3724/SP.J.1089.2024-00203
12. Tran, M. H., Gomez, O. & Fei, B. A video transformer network for thyroid cancer detection on hyperspectral histologic images. *InMedical Imaging 2023: Digit. Comput. Pathol.* **12471**, 32–41. https://doi.org/10.1117/12.2654851 (2023).
13. Jiang, P. et al. CSMViT: A lightweight transformer and CNN fusion network for lymph node pathological images diagnosis. *IEEE Access.* https://doi.org/10.1109/ACCESS.2024.3483769 (2024).
14. Hinton, G. Distilling the Knowledge in a Neural Network. Preprint at (2015). https://arxiv.org/abs/1503.02531
15. Lindqvist, J. et al. A general framework for ensemble distribution distillation. IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). 1–6, (2020). https://doi.org/10.1109/MLSP49062.2020.9231703 (2020).
16. Mirzadeh, S. I. et al. Improved knowledge distillation via teacher assistant. Proceedings of the AAAI conference on artificial intelligence. 34(04), 5191–5198 (2020).
17. Romero, A. et al. Fitnets: Hints for thin deep nets. Preprint at (2014). https://arxiv.org/abs/1412.6550
18. Ahn, S. et al. Variational information distillation for knowledge transfer. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 9163–9171 (2019).
19. Wang, H. et al. Progressive Blockwise Knowledge Distillation for Neural Network Acceleration. IJCAI. 2769–2775 (2018).
20. Chen, L. et al. Wasserstein contrastive representation distillation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 16296–16305 (2021).
21. Wang, L. & Yoon, K. J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (6), 3048–3068. https://doi.org/10.1109/TPAMI.2021.3055564 (2021).
22. Park, W. et al. Relational knowledge distillation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 3967–3976 (2019).
23. Shu, C. et al. Channel-wise knowledge distillation for dense prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision. 5311–5320 (2021).
24. Liu, Z. et al. A simple and generic framework for feature distillation via channel-wise transformation. Proceedings of the IEEE/CVF International Conference on Computer Vision. 1129–1138 (2023).
25. Kim, H. et al. RCKD: Response-Based Cross-Task knowledge distillation for pathological image analysis. *Bioengineering* **10** (11), 1279. https://doi.org/10.3390/bioengineering10111279 (2023).
26. Zhong, L. et al. Semi-supervised pathological image segmentation via cross distillation of multiple attentions and Seg-CAM consistency. *Pattern Recogn.* **152** https://doi.org/10.1016/j.patcog.2024.110492 (2024).
27. Li, L. et al. Global–local attention for image description. *IEEE Trans. Multimedia.* **20** (3), 726–737. https://doi.org/10.1109/TMM.2017.2751140 (2017).
28. Cibas, E. S. & Ali, S. Z. The 2017 Bethesda system for reporting thyroid cytopathology. *Thyroid* **27** (11), 1341–1346. https://doi.org/10.1089/thy.2017.0500 (2017).
29. Han Kai, X. et al. Transformer in transformer. *Adv. Neural. Inf. Process. Syst.* **34**, 5908–15919 (2021).
30. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 10012–10022 (2021).
31. Chen, C., Yu, J. & Ling, Q. Sparse attention block: aggregating contextual information for object detection. *Pattern Recogn.* **124**, 108418. https://doi.org/10.1016/j.patcog.2021.108418 (2022).
32. Huang, T. et al. Knowledge distillation from a stronger teacher. *Adv. Neural. Inf. Process. Syst.* **35**, 33716–33727 (2022).

Author contributions

Jiazhe Zhang and Fusong Jiang designed experiments; Haolin Zhang and Peng Jiang carried out experiments; Jiazhe Zhang and Haolin Zhang wrote manuscript; Qin Huang, Jingjing Chen and Yingling Cheng was responsible for the production and labeling of thyroid nodule cell pathology image dataset. Guangya Zhu analyzed experiments results; Shu Ran drew the manuscript, figures and tables.

Funding

This research is funded by Shanghai Research Center for Endocrine and Metabolic Diseases, grant number 2022ZZ01002; University of Shanghai for Science and Technology, grant number 2023JK-LY35Y.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

The original data used in this study have been granted with Ethics Approval No. YS-2024-268 from Shanghai Sixth People's Hospital affiliated to Shanghai Jiao Tong University, as required by the Institutional Review Board (IRB).

Additional information

Correspondence and requests for materials should be addressed to F.J.