Medical Image Analysis 89 (2023) 102916

Contents lists available at ScienceDirect

Medical Image Analysis

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

CReg-KD: Model refinement via confidence regularized knowledge distillation for brain imaging☆

Yanwu Yang [a,b], Xutao Guo [a,b], Chenfei Ye [b,d], Yang Xiang [b,**], Ting Ma [a,b,c,d,*]

[a] Electronic & Information Engineering School, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[b] Peng Cheng Laboratory, Shenzhen, China
[c] Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[d] International Research Institute for Artificial Intelligence, Harbin Institute of Technology (Shenzhen), Shenzhen, China

ARTICLE INFO

Keywords:
Knowledge distillation
Gating
Medical image
Regularization

ABSTRACT

One of the core challenges of deep learning in medical image analysis is data insufficiency, especially for 3D brain imaging, which may lead to model over-fitting and poor generalization. Regularization strategies such as knowledge distillation are powerful tools to mitigate the issue by penalizing predictive distributions and introducing additional knowledge to reinforce the training process. In this paper, we revisit knowledge distillation as a regularization paradigm by penalizing attentive output distributions and intermediate representations. In particular, we propose a Confidence Regularized Knowledge Distillation (CReg-KD) framework, which adaptively transfers knowledge for distillation in light of knowledge confidence. Two strategies are advocated to regularize the global and local dependencies between teacher and student knowledge. In detail, a gated distillation mechanism is proposed to soften the transferred knowledge globally by utilizing the teacher loss as a confidence score. Moreover, the intermediate representations are attentively and locally refined with key semantic context to mimic meaningful features. To demonstrate the superiority of our proposed framework, we evaluated the framework on two brain imaging analysis tasks (i.e. Alzheimer's Disease classification and brain age estimation based on T1-weighted MRI) on the Alzheimer's Disease Neuroimaging Initiative dataset including 902 subjects and a cohort of 3655 subjects from 4 public datasets. Extensive experimental results show that CReg-KD achieves consistent improvements over the baseline teacher model and outperforms other state-of-the-art knowledge distillation approaches, manifesting that CReg-KD as a powerful medical image analysis tool in terms of both promising prediction performance and generalizability.

1. Introduction

Advanced artificial intelligence technologies such as deep learning have shifted the research paradigm of medical image analysis (Isensee et al., 2021; Litjens et al., 2017). Prominent achievements of prior works manifest that deep learning approaches have become state-of-the-art solutions to a variety of medical image analysis problems, such as lesion detection (Albarqouni et al., 2016; Bejnordi et al., 2017; Dou et al., 2016), image segmentation (Baumgartner et al., 2017, 2019; Dou et al., 2020a), and phenotype prediction (Nandakumar et al., 2022; Venkataraman et al., 2011; Yang et al., 2021). These successes could be ascribed to effective learning algorithms and well-annotated medical data (Maier et al., 2017; Mehta et al., 2022; Menze et al., 2014). Nevertheless, it is still difficult to collect large, diverse, and well-annotated training sets in many cases (Hesamian et al., 2019; Razzak et al., 2018; Wong et al., 2018), whereas learning with insufficient samples might decrease the model performance in practice. Thus, it presents great challenges for deep learning approaches in medical image analysis to avoid over-fitting and poor generalizations.

Recently, a series of regularization strategies have been proposed to address this issue, such as label distribution learning (LDL) (Gao et al., 2017; Liao et al., 2020; Wang et al., 2022), label smoothing

☆ This work is done by Yanwu Yang during internship in Peng Cheng Laboratory.

* Corresponding author at: School of Electronics and Information Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen University Town, Rm 1206, Information Building, HIT Campus, Nanshan District, Shenzhen, Guangdong Province, China.
** Corresponding author.
E-mail addresses: 20b952019@stu.hit.edu.cn (Y. Yang), 18B952052@stu.hit.edu.cn (X. Guo), chenfei.ye@foxmail.com (C. Ye), xiangy@pcl.ac.cn (Y. Xiang), tma@hit.edu.cn (T. Ma).

https://doi.org/10.1016/j.media.2023.102916
Received 23 November 2022; Received in revised form 10 July 2023; Accepted 25 July 2023
Available online 26 July 2023

regularization (LSR) (Islam and Glocker, 2021; Müller et al., 2019) and multi-label learning (Mercan et al., 2017; Zhang and Zhou, 2013). These approaches soften the encoded targets to vectors of category distribution and penalize the output to improve generalization (Szegedy et al., 2016). Such labeling paradigms provide more knowledge of label relevance and can utilize dark knowledge, i.e. the knowledge of wrong predictions. However, it is still challenging to obtain good soft labels since they are difficult to be determined empirically and explicitly. A different approach to regularization is by knowledge distillation (KD) (Hinton et al., 2015), which was originally proposed to transfer knowledge from one model to another without a significant drop in performance. Apart from its usage on model compression, KD is feasible to penalize the predictive distributions and the intermediate representations with additional knowledge to improve generalization performance, e.g. from a pre-trained teacher model (Dou et al., 2020a; Rahimpour et al., 2021; Yang et al., 2021). It is worth noting that the knowledge distillation paradigm has been proven to be an adaptive version of label smoothing regularization (Chandrasegaran et al., 2022; Müller et al., 2019; Yuan et al., 2020), which prevents overconfident predictions and reduces intra-class variations. This provides more information than the simple one-hot encoding of the class labels and contains hidden inter-class dependency knowledge.

In particular, when the same network is used for both the teacher and student models, the paradigm is called self-knowledge distillation, which enables a model to learn knowledge from itself. In terms of this, self-knowledge distillation is known to be informative for the student network to regularize and refine its knowledge (Kim et al., 2021). Recently, this paradigm has been widely applied and has achieved promising improvements in various tasks in computer vision, such as image classification (Shen et al., 2022; Yun et al., 2020), semantic segmentation (Dou et al., 2020b; Ye et al., 2022), and brain age estimation (Yang et al., 2021). Especially, (Yuan et al., 2020) first proposes a Teacher-free Knowledge Distillation (TFKD) framework to refine a model itself by replacing the dark knowledge with predictions from the model. Progressive self-knowledge distillation (PSKD) gradually utilizes its knowledge for softening targets by training with a linear combination of the hard targets and past predictions (Kim et al., 2021). Apart from transferring knowledge to penalize the softmax output layer, FitNet (Romero et al., 2015) proposes to learn an intermediate representation that is predictive of the representations of the teacher. Feature refinement via self-knowledge distillation (FRSKD) further introduces an auxiliary self-teacher network to refine knowledge (Ji et al., 2021). Be your own teacher (BYOT) implements auxiliary classifiers to utilize the output of intermediate layers, where the knowledge in deeper networks is squeezed into shallow ones (Zhang et al., 2019). These approaches provide novel insights for enhancing the generalization performance and can be flexibly integrated into existing models.

However, most existing approaches transfer the entire knowledge without considering the global and local dependencies between the teacher and student. Firstly, the teacher model would not always bring good knowledge. If the teacher model mistakenly produces a wrong prediction with high probability, the student model would learn from the ground truth as well as a wrong imitation distribution, where the performance and prediction confidence might be decreased. Moreover, most existing approaches potentially ignore the complementary information between the teacher and the student representations. Reinforcing the student training straightforwardly with intermediate representations might bring in noise, and even lead to over-regularization (Song et al., 2022; Zhang et al., 2020). Finally, it is difficult to sufficiently utilize the feature distillation, since the high-dimensional intermediate features can be hardly leveraged to generate meaningful knowledge (Ji et al., 2021). To overcome these limitations, we propose to attentively transfer knowledge by considering knowledge confidence, which is conducted through a gating mechanism and attentive feature refinement layers.

In this regard, we revisit knowledge distillation as regularization paradigm for medical image analysis, especially high-dimensional 3D brain imaging with limited samples. This paper constitutes an extended version of our previous work (Yang et al., 2021), where we delve deeper into the investigation of global and local dependencies between teacher and student knowledge. To achieve this, we introduce the CReg-KD framework and provide a more comprehensive analysis through extended experiments and discussions. Building upon the Gated Distillation (GD) mechanism proposed in our previous work, which focuses on the global weighting of transferred knowledge, we propose an additional component called the Attentive Feature Refinement layer (AFR). This layer enables the refinement of feature map transformation by incorporating attentive local semantic context. By combining these elements, CReg-KD guides the student model to learn from both global and local dependencies. Our evaluation expands to encompass AD classification and brain age distribution learning tasks using T1w images. Extensive experimental results demonstrate the superiority of our proposed CReg-KD in medical image analysis by consistently improving performances over state-of-the-art methods and enhancing generalizability without increasing parameters for the student model. Our code is available at https://github.com/podismine/CReg-KD. The main contributions of the proposed work are:

- We introduce a novel self-supervised knowledge distillation framework called CReg-KD, which serves as a regularization paradigm to enhance performance and generalizability.
- Global and local dependencies are investigated to regularize the transfer of knowledge by the gating and attentive refinement layers.
- Through extensive experiments on two tasks, we demonstrate the superiority of our method in improving performance and generalization.

In this extended version, we introduce the following enhancements:

- Characterizing the confidence of transferred knowledge based on global and local dependencies, resulting in the development of the CReg-KD framework. Specially, we propose the AFR layers to refine the feature map using attentive local semantic context.
- A comparison of methods on an additional task: AD classification based on T1w images. We conduct extended evaluations, including ablation studies, model generalizability, and performance analysis with different sample sizes. We also provide a sensitivity analysis of various types of AFR.
- An improved performance in brain age prediction and outperforming results for AD classification. Notably, even with a simple baseline like ResNet-18, our CReg-KD approach is able to achieve top-performing results.

2. Methodology

In this section, we first overview the knowledge distillation techniques in Section 2.1 and then introduce the proposed gated mechanism for knowledge distillation in Section 2.2.1. Moreover, attentive feature refinement layers for attentive intermediate representations learning are investigated in Section 2.2.2.

2.1. Preliminate of knowledge distillation

Knowledge distillation was first proposed to transfer knowledge, usually from a larger deep neural network into a small network without a significant drop in performance. The main idea behind this is that the student model mimics the teacher model to obtain a competitive or even superior performance. In self-distillation, the teacher and student model come from the same neural network, and the student utilizes its knowledge to improve itself.

Formally, we denote T and S as the teacher and student networks respectively. The teacher model is firstly pre-trained. And the student

model learns from a weighted combination of the ground truth and the output distribution of the teacher:

$$\mathrm{L} = \alpha \mathrm{L}_{\mathrm{gt}} + (1-\alpha)\cdot \mathrm{L}_{\mathrm{KL}} \tag{1}$$

$$\mathrm{L}_{\mathrm{KL}} = \mathrm{KL}(\mathrm{p}^{\mathrm{T}} \parallel \mathrm{p}^{\mathrm{S}}) \tag{2}$$

where KL denotes the Kullback-Leibler (KL) divergence, which is used to measure the distribution similarity between the teacher output p^T and the student output p^S. The objective function is combined with the target loss L_{gt} and the imitation loss L_{KL} raised by the knowledge distillation with a hyperparameter α to balance two terms. Moreover, a relaxation temperature τ is usually introduced to soften the signal arising from the teacher output:

$$p^T(x;\tau) = softmax\left(\frac{z^T(x)}{\tau}\right) \tag{3}$$

$$p^S(x;\tau) = softmax\left(\frac{z^S(x)}{\tau}\right) \tag{4}$$

Finally, the formulation for the student training with KD can be obtained as a regularization form by multiplying a square of the temperature τ^2 to stabilize the back-prop gradient:

$$L = \alpha L_{gt} + (1-\alpha)\tau^2 \cdot KL\left(p^T(x;\tau) \parallel p^S(x;\tau)\right) \tag{5}$$

Notably, the first term in Eq. (5) is the original loss between the student output and the ground truth, and the second term reinforces the student network to learn from the softened output of the teacher model. And thus self-distillation learning can be viewed as learning from the target and a regularization term.

In particular, when it comes to the classification task, as an example, Eq. (5) is represented as:

$$L_{KD} = \alpha L_{CE}\left(p^{GT},\ p\right) + (1-\alpha)\tau^2 \cdot KL\left(p^T(x;\tau) \parallel p^S(x;\tau)\right) \tag{6}$$

If we formulate the smoothed label distribution as $q(x) = \alpha p^{GT}(x) + (1-\alpha)u(x)$ with $u(x)$ as a uniform distribution, the label smoothing regularization can be formulated as:

$$L_{LSR} = \alpha L_{CE}\left(p^{GT}, p\right) + (1-\alpha)\left(KL(u(x) \parallel p(x) + H(u))\right) \tag{7}$$

, where H denotes the entropy. If we set the temperature $\tau = 1$, then we obtain $p^S(x) = p(x)$. Eq. (6) can be viewed as a special case of Eq. (7) with a learned distribution from the teacher $p^T(\mathrm{x})$. In this regard, the knowledge distillation paradigm is a special case of LSR and can be utilized as a regularization paradigm.

2.2. Confidence regularized knowledge distillation

As stated above, the knowledge distillation paradigm provides a learned and adaptive label regularization to reinforce student training. Based on this mechanism, we take the dependencies between teacher and student knowledge into consideration. Fig. 1 demonstrates the structure of our proposed CReg-KD framework including gated distillation, output probability distillation, and attentive feature refinement. The gated distillation penalizes the transferred knowledge globally by weighting the teacher loss, while the attentive feature refinement layers provide locally enhanced features with key semantic contexts.

2.2.1. Gated distillation

Gating is a popular strategy for controlling information passing. Long short-term memory and gated recurrent unit are well-known

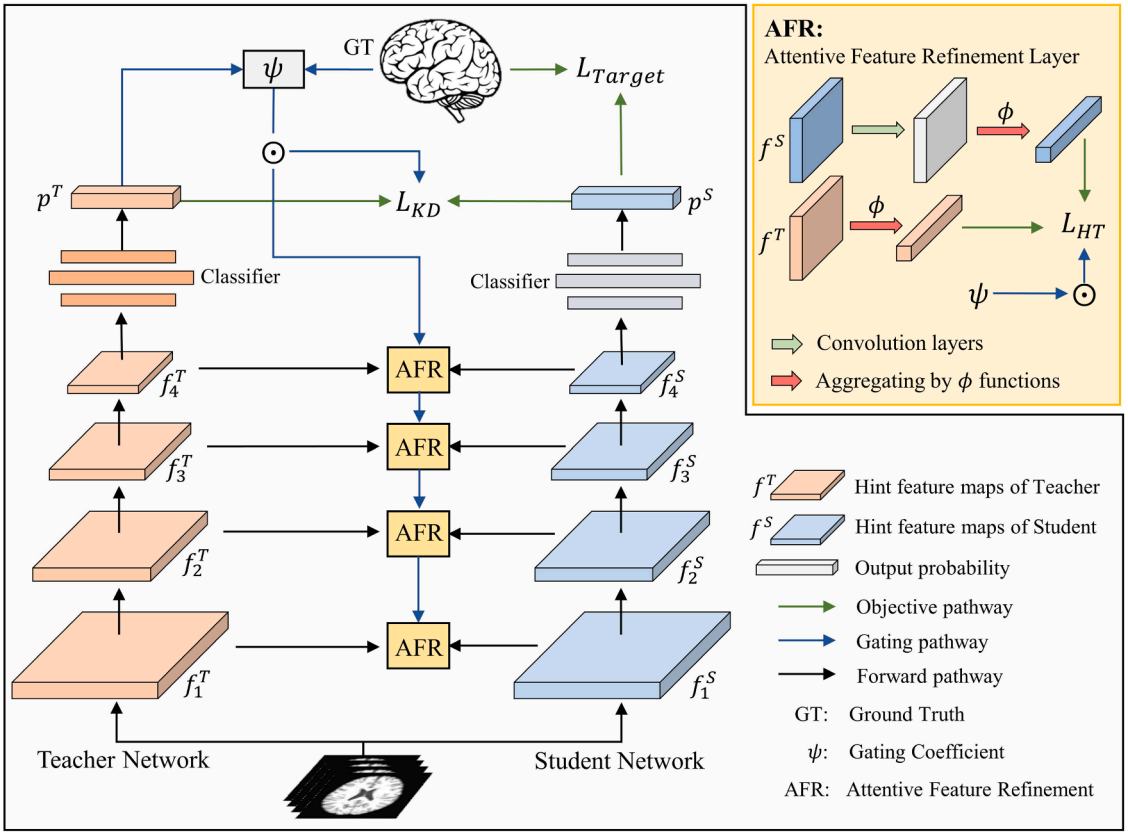

Fig. 1. Overview of the proposed Confidence Regularized Knowledge Distillation (CReg-KD) framework, where the output probability and attentively refined feature maps are gated to transfer confidence-regularized knowledge. The details of Attentive Feature Refinement (AFR) layers are demonstrated in the upper right of the figure, where the convolution layers are built by two convolution operators with a kernel size of 3 and 1 respectively followed by a ReLU activation function. The objective loss is obtained by combining the task target loss L_{target}, output distillation loss L_{KD}, and refined feature distillation loss L_{HT}.

architectures that leverage gating memory for dealing with context-sensitive representations, and achieve impressive results on sequence modeling (Chung et al., 2014; Hochreiter and Schmidhuber, 1997; Peters et al., 2017). In this section, we introduce a novel gated distillation mechanism for filtering knowledge transferred.

In the knowledge distillation paradigm, the student model learns from two aspects: the target and the knowledge of the teacher model. Considering that the teacher model does not always provide meaningful knowledge, it is problematic to guide the student model to learn from two inconsistent targets. In this regard, we regularize the student model to learn from the teacher model when the teacher provides confident information. In detail, we re-weighted the teacher prediction error as a confidence score to guide the student model for training. Intuitively, when the error is too large to supervise, the student would only learn from the ground truth by itself without learning from the teacher model. The Gated Distillation (GD) mechanism is obtained as:

$$\psi_k = 1 - clip\left(\frac{dis\left(o_k^T, o_k^{GT}\right)}{\eta}\right) \tag{8}$$

where ψ_k denotes the transferring weight of the k-th sample, η rescales the prediction error, and *clip* function restraints the weights to $\psi_k \in (0, 1)$. o_k^T and o_k^{GT} denote the outputs of the teacher and the target respectively. Function *dis* represents a distance function to measure the similarity of the teacher output and the target. This function is decided in the experiments according to different task objectives. The setting of η and *clip* operation limits the prediction error with an upper bound. We consider that the teacher model could not give meaningful knowledge to the student model when the prediction is out of this range. Our implementation helps to mitigate the disturbance of the unconfident transferred knowledge and contributes to more stable training. With the gated distillation mechanism, we would obtain the objective function with ψ as:

$$L = \alpha L_{target} + (1-\alpha)\tau^2 \cdot \psi KL\left(p^T(x;\tau) \parallel p^S(x;\tau)\right) \tag{9}$$

2.2.2. Representation refinement with attentive semantic context

Intermediate representations are predictive of the teacher network and allow for generalizing without utilizing soft targets (Romero et al., 2015). Although the hint representations convey more meaningful knowledge compared to the soft targets, the high dimensionality of the intermediate feature maps might hinder the imitation of knowledge transfer, which decreases the performances. In this regard, it is problematic to simply transfer the entire high-dimensional hint representations for knowledge transfer. The gated distillation mechanism mentioned above allows for filtering sample-wise knowledge, however, is limited to hint representations. In this section, we introduce an auxiliary structure to generate refined hint representation by incorporating sample knowledge filtering and hint representation refinement for mimicking the generalization capability of the teacher. The idea behind this is inspired by previous studies that leverage auxiliary convolution blocks for multi-scale or cross-level feature alignment (Qi et al., 2021). We hypothesize that the auxiliary learning architecture could also refine and develop key hint representations for imitation.

Apart from the soft target distillation, feature distillation induces the network to mimic the refined hint representations:

$$L_{HT} = \left|\left|\phi\left(\widehat{h}_l^S\right), \phi\left(h_l^T\right)\right|\right|^2, \widehat{h}_l^S(x) = f\left(h_l^S\right) \tag{10}$$

where the hint representations of the student are refined by the f operation, which is built by two convolution layers with a kernel size of $[3, 1]$ respectively. L_2 loss is implemented to minimize the discrepancy between the representations h_l of the l-th layer. ϕ is a pooling function for aggregating spatial attention maps: $\phi : R^{C\times H\times W\times D} \rightarrow R^{H\times W\times D}$. More specifically, we will consider the following spatial attention maps:

- Averaged absolute values raised to the power: $\phi_{pow-avg}(x) = \sum_{i=1}^{C} \frac{|x_i|^2}{C}$.
- The power of the averaged values: $\phi_{avg-pow}(x) = \left(\sum_{i=1}^{C} \frac{x_i}{C}\right)^2$.
- The average of the values: $\phi_{mean}(x) = \sum_{i=1}^{C} \frac{x_i}{C}$.
- The raw features: $\phi_{raw}(x) = x$.

Moreover, we extract the feature maps of M layers for distillation to minimize the discrepancy between teacher and student representations. In this study, feature maps of $M = 4$ layers are refined and averaged. Finally, by integrating the loss functions above, the objective function is formulated as:

$$\begin{aligned} L &= \alpha L_{target} + (1-\alpha)\psi L_{KD} + (1-\alpha)\psi L_{HT} \\ &= \alpha L_{target} + (1-\alpha)\tau^2 \psi KL\left(p^T||p^S\right) + (1-\alpha)\psi \sum_{l=1}^{M} \left|\left|\phi\left(\widehat{h}_l^S\right), \phi\left(h_l^T\right)\right|\right|^2 \end{aligned} \tag{11}$$

Moreover, to better balance the loss values of the three terms, we reformulate the loss function with the weights λ_1 and λ_2:

$$L = L_{target} + \lambda_1 \tau^2 \psi KL\left(p^T||p^S\right) + \lambda_2 \psi \sum_{l=1}^{M} \left|\left|\phi\left(\widehat{h}_l^S\right), \phi\left(h_l^T\right)\right|\right|^2 \tag{12}$$

3. Experiments

3.1. Materials

In this study, to evaluate the robustness and effectiveness of the proposed method, we made comparisons on two tasks including the classification of Alzheimer's disease (AD) and brain age prediction. The descriptive information is displayed in Table 1.

AD classification. The AD classification task is carried out to distinguish AD from healthy participants. We leveraged the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1] to form the cohort, including a total of 902 participants. Among them, 439 participants were diagnosed as AD at baseline and 463 participants were healthy at their first examination. These subjects are divided into two groups: AD and healthy control (HC) in accordance with the standard clinical criteria, such as Mini-Mental State Examination (MMSE) scores and Clinical Dementia Rating (CDR). The sex and age of the HC and AD group are matched.

Brain age prediction. Previous studies demonstrated that MRIs could be used to predict chronological age and show that brain age is vital to help improve the detection of early-age neurodegeneration and predict age-related cognitive decline (Bashyam et al., 2020; Jónsson et al., 2019; Ran et al., 2022). Accurate brain age estimating is an essential prerequisite for quantifying the predicted age difference as a

Table 1
Demographic details on two tasks including gender, age, Mini-mental State Examination (MMSE), and Clinical Dementia Rating (CDR).

Task	Group type	Gender (Male/ Female)	Age (Mean ±Std)	MMSE (Mean ±Std)	CDR (Mean ±Std)
AD classification	HC	258/205	73.93 ±6.82	28.93 ±1.43	0.00 ±0.00
	AD	240/199	75.03 ±7.87	21.80 ±3.74	0.77 ±0.27
Brain age estimation	HC	1569/ 2086	57.31 ±21.53	–	–

[1] http://adni.loni.usc.edu

biomarker. In this study, a total of 3655 T1-weighted MRI images from four public datasets including the IXI database,[2] the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Open Access Series of Imaging Studies (OASIS) (Marcus et al., 2010), and 1000 Functional Connectomes Project[3] (1000-FCP) are selected to form our cohort. Only healthy subjects were selected in our experiments, with no indication of neurological pathology, and no psychiatric diagnosis. The ADNI and OASIS datasets are public with longitude studies, where 1024 and 1028 adult and elderly subjects with ages ranging from 42 to 92 are included. The 1000-FCP projects mainly cover the young with 1040 subjects with a mean age of 25, and the IXI dataset covers 563 subjects with a full range of ages.

Preprocessing. All the images were acquired at 3T or 1.5T T1-weighted MRI. The images were processed including AC-PC aligns, brain skull stripping, bias field correction (Sled et al., 1998), and linear normalization into the standard MNI space. Additionally, z-score normalization is employed to narrow the gap between different data centers and is vital for successful deep learning-based MR image synthesis (Reinhold et al., 2019). After preprocessing, all images are down-sampled trilinearly into the standard $2mm^3$ MNI space and padded into the size of $96 \times 112 \times 96$.

3.2. Competitive methods

To test our proposed method, we compare it with several state-of-the-art knowledge distillation approaches. These methods include Teacher-free knowledge distillation (TFKD) (Yuan et al., 2020), Progressive self-knowledge distillation (PSKD) (Kim et al., 2021), FitNet (Romero et al., 2015), Feature Refinement via Self-Knowledge Distillation (FRSKD) (Ji et al., 2021) and Be your own teacher (BYOT) (Zhang et al., 2019), which are mentioned above. We follow the training strategy by reference to the original architecture and implement the temperature within a grid search of $[0.5,0.9,1,3,5,10]$. The parameter α_T at the last epoch in PSKD is set as 0.8. For FitNet, the layer to transfer intermediate representations is searched and decided according to different architectures. The auxiliary architectures in BYOT and FRSKD are modified by 3D convolution groups. We evaluate the effect of knowledge distillation in penalizing the softmax output by comparison with TFKD and PSKD. FitNet, FRSKD, and BYOT are used to test the knowledge distillation by feature learning.

3.3. Implementations

Network architecture. For each task, four well-estimated neural network frameworks are implemented for measuring. (1) AD classification. The ResNet with 18/50 layers (He et al., 2016), DenseNet121 (Huang et al., 2018), and Inception-V3 (Szegedy et al., 2016) are generally purposed models for classification. These four models were used as the backbone for feature learning. For each network, the default 2D operations were replaced by 3D, and a global average pool layer was applied to average the encoded features. Finally, a two-layer multiple-layer perception was used for classification. The distance function *dis* in Eq. (8) for the AD classification is $dis(o_k^T, o_k^{GT}) = |o_k^T - o_k^{GT}|$, where $o_k^T = p_k^T$, $o_k^{GT} = p_k^{GT}$. (2) Brain age prediction. We employed two well-estimated models including ResNet18, and ResNet50 and two models that are specially designed for brain age prediction: SFCN (Peng et al., 2021), and DeepBrainNet (Bashyam et al., 2020). To fit our 3D neuroimaging data, we utilized the standard architecture and replaced the 2D operations with 3D. And a global average pool was applied to average the features. These models encode the image data into 88 features. A multiple-layer perception with three layers was utilized to classify the features into 22 probabilities followed by a softmax function. We leveraged the label distribution learning paradigm by converting the single age value into a normal distribution, which is a well-known strategy for age regression (Gao et al., 2017; Peng et al., 2021). The variance of the normal distribution was set as 2. The distance function is implemented as $dis(o_k^T, o_k^{GT}) = \left| \sum_m^{22} l_m q_m^T - \sum_m^{22} l_m \widehat{q}_m \right|$, where $\widehat{q}_m$ and q_m^T denote the label distribution of ground truth and the prediction of the teacher, and l_m denotes the m-th element in the label set $L = (l_m = 12 + 4k | m = 0,1,\ldots,21)$. More details could be found in our previous studies (Yang et al., 2021).

Optimization. Our experiments on two tasks were carried out on 4 NVIDIA 2080Ti GPUs with 11GB memory. The networks were trained by the Adam optimizer on the PyTorch 1.6 platform. Moreover, cross-entropy was applied as the loss function for the AD classification task. For brain age estimation, a KL divergence was used to measure the similarity between the output prediction and the manually designed distribution.

Training details. The networks were trained with an initial learning rate of 1e−6, and an L2 wt decay coefficient of 5e−5. The learning rate was increased linearly to 1e−4 in 10 warmup epochs. The best model was obtained based on the validation loss. The models were trained in 120/260 epochs for the AD classification and brain age prediction respectively. The batch size was set to 32. To reduce the risk of overfitting, two data argumentation methods were applied during training: random rotation and random shifting. The rotation angles were between -10° and 10°, and the input was randomly shifted by 5 voxels along every axis with equal probability. Hyper-parameters of temperature τ and gating scale η would be discussed further in the results.

Measurement. Moreover, to evaluate the performance of competitive methods, different metrics were implemented for various tasks. For the AD classification task, the accuracy (ACC), Sensitivity (SEN), Specificity (SPE), and area under the curve (ROC-AUC) were used for measuring the performance of classification. Mean absolute error (MAE), Pearson correlation coefficient (PCC), and cumulative score (CS) between the predicted ages and chronological ages were used for the brain age prediction task. PCC measures the correlation between the predicted ages and the chronological ages. The CS is the accuracy of age estimation within a threshold α, which is obtained by: $CS(\alpha) = \frac{N_{e \leq \alpha}}{N} \times 100\%$, where $N_{e \leq \alpha}$ is the number of samples on which the absolute error of prediction e is no higher than the threshold α. For these metrics, a higher value indicates a better performance. In addition, we implemented the Expected Calibration Error (ECE) and negative log loss (NLL) for measuring whether the predictions were well calibrated, approximating the difference in expectation between classification accuracy and confidence estimates. The predictions were partitioned into 10 bins and taken as a weighted average of bins' difference of confidence and accuracy to evaluate whether the model represents the true correctness likelihood. For ECE and NLL, a lower value is better. For each task, the experiments were repeated four times, where the samples are randomly sampled into a train set (90%), and a validation set (10%). We fixed the validation set as well as the random seed for each fold and averaged the results for comparison.

4. Results

In this study, we evaluated our proposed self-distillation approach on two tasks. In the following sections, we first demonstrate details of the parameter setting in Section 3.1 and then show results of the applications on the AD classification and brain age prediction tasks in Section 3.2 and Section 3.3 respectively.

4.1. Parameter discussion

In our study, we tuned the hyperparameters to check the effect of the

[2] http://brain-development.org
[3] http://www.nitrc.org/projects/fcon_1000

performance. Considering that a full grid search for all these hyper-parameters costs numerous time and resources, we first fixed the relaxation temperature τ and searched within [0.5, 0.9, 1, 3, 10] for both tasks. With obtained relaxation temperature, we searched the loss weights λ_1 and λ_2 to measure the balancing of the combined loss within the set of [0.2, 0.4, 0.6, 0.8, 1.0].

These parameters of the best models are listed in Table 2. For the AD classification task, all four models achieve the best performance with a temperature of 0.9. While the loss weights λ_1 and λ_2 are sensitive to performance, and are critical parameters that determine the extent of probability and feature transferring context for regularization. One might wonder if the architectures are related to these choices. For simple architectures like ResNet-18, it is available to use higher values for the loss weights for strong regularization. However, for more complex architectures such as InceptionV3, higher values could lead to over-regularization and decrease performance. In the brain age prediction task, the temperatures of 3/10/3/10 are the best for the ResNet-18, ResNet-50, SFCN, and DeepBrainNet respectively. Low temperature settings are known to protect models from noisy negative labels, while high temperature settings mitigate the peakiness of teacher logits and enable learning from negative labels (Cho and Hariharan, 2019; Zhou et al., 2021). For the brain age distribution learning, a higher temperature contributes to capturing the label ambiguity between ages. However, it is important to note that the performance does not show significant changes with different temperature settings. Besides, regarding the loss weights, they are set to be the same since they only have a slight impact on performance in the brain age distribution learning task. To summarize, a temperature setting of 0.9 is effective for achieving high performance in most classification cases, while a higher temperature value is preferable for distribution learning. The choice of $\lambda_1 = 0.8$ is suitable for the majority of scenarios.

4.2. Results of AD classification

4.2.1. Comparison with competitive methods

Table 3 shows the validation results on four well-estimated CNN models including ResNet-18, ResNet-50, Densenet-121, and InceptionV3. The performances of accuracy, sensitivity, specificity, and ROC-AUC are listed, where the best results are marked in bold, and the second best are underlined. Statistical results are displayed to measure the improvements, where * denotes the performance is significantly improved with a p-value < 0.05. The baseline column indicates the teacher model. From these results, we can see that distillation on both the softmax output (e.g. TFKD, PSKD) and the intermediate features (e.g. FitNets and FRSKD) achieve improvements on four well-estimated baseline models. However, not all the distillation methods achieve consistent improvements in all four architectures. For example, FRSKD achieves comparable accuracy results with the baseline teacher method by using ResNet18 (ACC: 88.05% vs. 87.05%) and InceptionV3 (ACC: 90.21% vs. 90.21%). Overall, PSKD and TFKD achieve consistent improvements over the baseline teacher model and outperform other approaches in most architectures.

Compared with these state-of-the-art models, our proposed CReg-KD achieves consistent improvements on all four architectures. Especially, CReg-KD outperforms other state-of-the-art distillation approaches with 2.8%, 0.41%, 0.4%, and 1.4% improvements. Moreover, ResNet-18 using CReg-KD enhances the accuracy from 87.05% to 92.35% without increasing model parameters in the student model. Apart from the promising results, we can also see that the distillation approaches could enhance the performance without limitation to the baseline models, where improvements are achieved on all four models. The ROC curve comparison is shown in Supplementary A.1, where the curves are obtained by validation results on all folds and then smoothed. Overall, the InceptionV3 model with CReg-KD achieves the best performance with an accuracy of 94.05%, a sensitivity of 94.65%, a specificity of 93.57%, and a ROC-AUC of 94.11%.

Table 2
The hyper-parameter settings for the four models in two tasks including the temperature τ, loss weights λ_1 and λ_2.

Task	Model	τ	λ_1	λ_2
AD classification	ResNet-18	0.9	1.0	1.0
	ResNet-50	0.9	0.8	0.2
	DenseNet-121	0.9	0.8	0.2
	InceptionV3	0.9	0.4	0.6
Brain age estimation	ResNet-18	3.0	0.8	0.8
	ResNet-50	10.0	0.8	0.8
	SFCN	3.0	0.8	0.8
	DeepBrainNet	10.0	0.8	0.8

4.2.2. Generalization performance

As shown in Fig. 2, the learning curves (upper) and the differences (lower) in terms of accuracy are plotted on four well-estimated models. In the learning curves, the accuracies of training and validation in the teacher model are shown in dark and light blue, and those in the student are figured in dark and light orange. From the results, we can observe that the student model gets convergence faster than the teacher model (blue vs. orange) in both training and validation. It is reasonable as the knowledge distillation facilitates the student to learn from additional knowledge provided by the teacher model; The knowledge distillation paradigm penalizes the student learning with calibrated knowledge, which optimizes the backpropagation with regularized gradient. In addition, the differences between the training and validation accuracy are plotted below the learning curves, where the blue and orange indicate the differences in the student and teacher models respectively. The differences of student models on all four architectures tend to be smaller than the teacher models, which indicates that the overfitting is mitigated and generalization performances are enhanced to some extent. Especially on the ResNet-18 model, the difference between training and validation loss of the student model is significantly reduced in the beginning process, compared with the teacher model. This might lead to the model getting convergence more stably and robustly.

In addition, we listed the expected calibrated error rate and the negative log loss in Table 4. The two measurements are used to evaluate the quality of predictive probabilities in terms of confidence estimation. The results in the table show that our proposed CReg-KD achieves better performance on confidence estimation. Compared to the baseline teacher model, 1.08%, 0.95%, 0.44%, and 0.53% ECE, 0.12, 0.03, 0.1, and 0.03 NLL are reduced on ResNet-18, ResNet-50, DenseNet-121, and InceptionV3 respectively. Overall, CReg-KD achieved the best confidence estimation performance on most architectures.

4.2.3. Ablation study

Moreover, in this study, we perform ablation studies to measure the effect of each component involved in CReg-KD. Table 5 listed the results of the studies on the effect of the gating mechanism, softmax output distillation, and attentive feature refinement, where the results in the first 2 rows are the same as those of the teacher model and TFKD. Moreover, from the results, we can see that as the number of components increases, the performances are enhanced progressively. In particular, the gating mechanism plays a key role in improving performance, which further enhances accuracy. For example, with the gating mechanism, the accuracy is increased from 87.75% to 90.93% on the ResNet-18 model even without feature distillation. However, the performance of DenseNet-121 dropped by 0.81% with the gating (row2: 91.96% vs. row5: 91.15%). This might be caused by an unsuitable gating, which might regularize the student learning strictly and cause over-regularization. Nevertheless, by gating the softmax outputs and features, DenseNet-121 outperforms other ways of distillation. In addition, the second-to-last line demonstrates the performance of our previous method (Yang et al., 2021) which utilizes the gating mechanism and raw feature distillation for regularization. Comparing these results to our

Table 3
Comparison results on AD classification in terms of accuracy (Acc%), sensitivity (Sen%), specificity (Spe%), and the area under the curve (ROC-AUC%). Competitive approaches TFKD, PSKD, FitNets, FRSKD, and BYOT are included for evaluation. The best results are shown in bold, and the second best are shown with an underline with average and standard deviation across folds (Mean±Std).

Models	Metric	Baseline	TFKD	PSKD	FitNets	FRSKD	BYOT	CReg-KD
ResNet-18	ACC↑	87.05±5.00	87.75±4.38	90.55±2.11*	87.39±4.02	88.05±3.36	89.87±2.23	**92.35±2.10***
	SEN↑	91.32±4.55	82.56±7.14*	92.02±3.14	88.40±5.80	84.10±7.72	87.00±5.83	**93.94±2.36**
	SPE↑	83.11±6.99	**92.57±2.24***	89.19±3.31*	86.49±5.41	89.86±4.83*	**92.57±2.24***	91.89±3.31*
	ROC-AUC↑	87.22±4.95	87.57±4.47	90.60±2.08*	87.44±4.02	86.98±3.46	89.78±2.31	**92.41±2.06***
ResNet-50	ACC↑	88.45±1.58	89.51±2.51	91.61±1.00*	88.81±2.24	90.90±1.62	90.55±2.31	**92.02±2.48***
	SEN↑	87.95±1.46	90.59±3.72	92.06±3.1	86.28±5.40	92.73±2.55	92.04±2.35	**94.94±3.77***
	SPE↑	89.86±2.24	88.51±3.51	**91.22±2.24**	**91.22±6.16**	89.19±1.91	89.19±3.31	88.51±5.85
	ROC-AUC↑	88.41±1.57	89.55±2.50	91.64±1.04	88.75±2.18	90.96±1.62	90.61±2.29	**91.73±1.89**
DenseNet-121	ACC↑	86.01±2.22	91.96±1.18*	88.80±2.29	89.50±3.54	88.46±3.33	91.95±1.21*	**92.36±2.27***
	SEN↑	89.81±6.11	93.47±2.42	87.63±4.43	89.83±3.31	89.81±3.37	92.00±4.37	**94.20±2.90**
	SPE↑	82.43±4.05	90.54±1.35*	89.86±2.24	89.19±6.62*	87.16±7.25	**91.89±4.27***	89.19±5.06
	ROC-AUC↑	86.12±2.32	92.00±1.20*	88.75±2.34	89.51±3.44	88.49±3.24	91.94±1.20	**92.70±2.19***
InceptionV3	ACC↑	90.21±3.11	92.65±1.17*	91.60±2.02	92.65±2.09	90.21±1.97	92.31±1.56	**94.05±1.20***
	SEN↑	89.71±7.10	93.47±3.23	93.47±3.81	92.75±4.35	89.42±6.15	94.20±0.08	**94.65±1.45**
	SPE↑	92.57±5.19	91.89±2.70	89.86±2.24	92.57±5.19	91.89±5.06	90.54±3.02	**93.57±2.24**
	ROC-AUC↑	90.14±3.14	92.68±1.20	91.67±2.05	92.66±2.03	90.16±2.00	92.37±1.51	**94.11±1.16**

* : significant outperforming with the p-value < 0.05.

proposed feature refined layers, we observe improved performances across all four architectures consistently. This improvement can be attributed to the integration of attentive local contexts, which adaptively supervise the learning of the student model. Overall, with the gating of both softmax output and feature representations, all four well-estimated CNN models achieve the best performances.

Moreover, we figured out the effect of the gating scale setting in Fig. 3. It is shown that as the scale increases, the performances in terms of accuracy tend to increase and then decrease. The optimal results are achieved with a scale of 0.3 or 0.35. In addition, the performances are comparable to those of TFKD or FRSKD when the scale is greater than 0.35, which indicates that most knowledge is transferred for distillation. The detailed values are listed in Supplementary materials A.2. Furthermore, from Table 3 and Fig. 3, we can see that an unsuitable setting of scale might decrease the performances, where the performances with the gating are lower than those without the gating. For example, the DenseNet-121 without the gating (91.25%) achieves better than that using the gating with a scale lower than 0.2 (0.05: 90.90%; 0.1: 90.90%; 0.15: 90.56%). We suspect that the teacher model might provide over-regularization information to regularize student learning.

Furthermore, the effects of the way for feature refinement are measured and plotted in Fig. 4 with error bars and averaged values of accuracy performance. It can be seen that refining intermediate representations by power and mean ϕ_{pow_mean} outperforms other feature aggregation methods. And consistent improvements are obtained in all four architectures. Especially, compared with using raw features, 2.21%, 0.76%, 2.84%, and 2.05% improvements are obtained on ResNet-18, ResNet-50, DenseNet-121, and InceptionV3 respectively. Besides, refining by ϕ_{mean} decreases the performance significantly in most models, and the performances are even lower than using the raw features. We suspect that the simple averaging operation might harm the injective transformation of representations, and limit the spatial diversity of the feature maps.

4.2.4. Evaluation on different sizes of samples

To further evaluate the performance and generalizability of our proposed CReg-KD method on medical images with limited samples, we conducted comparisons by training models with different sample sizes (i.e., 25%, 50%, 75%, and 100% of the training samples). It's worth noting that the testing samples remained fixed and consistent with previous comparisons. Fig. 5 presents the accuracy performance of four architectures using different regularization approaches: baseline (red), TFKD (orange), PSKD (green), BYOT (black), and our CReg-KD (blue). These regularization methods demonstrate superior performance in most cases for both AD classification and brain age prediction tasks. The results show a consistent drop in performance for the baseline models as the training samples decrease across all four architectures. Additionally, architectures with more parameters, such as InceptionV3, outperform other models with varying sample sizes. Moreover, our proposed CReg-KD method stands out as the top-performing regularization approach among all the knowledge distillation methods. Notably, InceptionV3 with CReg-KD, using only 75% of the data (Acc: 93.76%), achieves comparable performance to using 100% of the data (Acc: 94.05%). This can be attributed to both the effective model architecture of InceptionV3 and the proposed knowledge distillation term, which enhances generalizability. Furthermore, our CReg-KD significantly improves the performance of ResNet-18 and ResNet-50 models with only 25% training data, with improvements of 4.75% and 5.37% respectively. This indicates that our CReg-KD is a promising tool for enhancing performance and generalization in cases of limited medical images. Furthermore, we suspect that even training with 100% data may not be sufficient to overcome overfitting and poor generalization, as the number of involved samples still remains insufficient. Therefore, the observed improvements overall can also be attributed to addressing the generalizations. In summary, our findings demonstrate the effectiveness of CReg-KD in improving performance and generalization with limited medical image samples, and highlight its potential for addressing the challenges posed by insufficient data in medical imaging tasks.

4.3. Results of brain age prediction

4.3.1. Comparison with competitive methods

To further verify our proposed CReg-KD for medical image analysis, we also evaluated its application on brain age prediction. The experimental results are shown in Table 6, where the averaged results and standard deviation across folds in terms of MAE, PCC, and CS scores are listed. The statistical analysis was performed by the Wilcoxon test, where significant improvements with a p-value below 0.05 are shown with * in the table. From the results, we can see that the two specially designed models, i.e. SFCN and DeepBrainNet, outperform the general-purpose models (ResNet-18 and ResNet-50). Especially, the SFCN model with fewer parameters could obtain promising performances for brain age estimation, which corresponds to previous findings (Peng et al., 2021). Moreover, CReg-KD performs the best on all four well-estimated models and achieves consistent improvements over other regularization methods, where the DeepBrainNet model achieves the best performance with the MAE of 2.162, the CS of 90.986%, and the PCC of 0.989. The scatter plots of predictions are displayed in A.3 in the supplementary

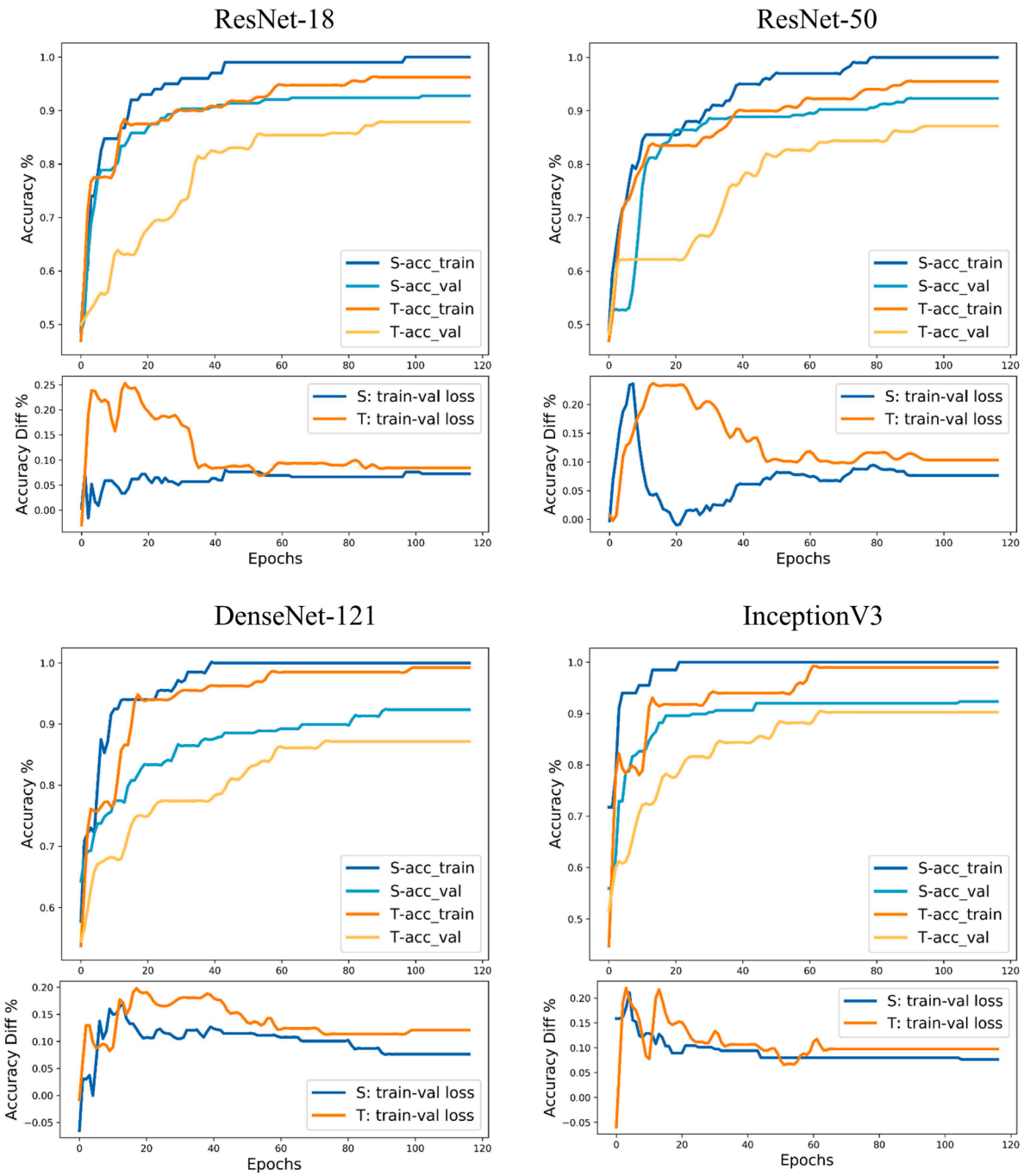

Fig. 2. Plots on the learning curves and differences of four architectures in terms of accuracy, where the performances of ResNet-18, ResNet-50, DenseNet-121, and InceptionV3 are figured. The learning curves were smoothed by plotting the best performance before the current epoch. The orange and the blue represent the teacher model and the student model respectively. The differences in accuracy between the training and validation are shown below the learning curves. The student models tend to get convergence faster than the teacher models and achieve smaller differences.

materials. Both the experimental results and statistical analysis indicate that our proposed method is a promising and powerful tool for model regularization to improve performance.

4.3.2. *Generalization performance*

On the other hand, we demonstrate the error rate results in Table 8, where the KL divergence score is applied to measure the errors between the learned outputs and the target distributions. The standard deviation values are ignored since the values are too small for comparison. The results demonstrate that our proposed CReg-KD framework achieves the lowest error in all four architectures, indicating that the generalization ability is better than other methods. Moreover, the learning curves and differences are plotted in Supplementary A.4. The results correspond to those in the AD classification task, where the student models achieve smaller differences than the teacher model in most architectures. Overall, our proposed CReg-KD is robust and generalizes to brain age estimation.

4.3.3. *Ablation studies*

Finally, we applied ablation studies for the brain age estimation task. Table 7 lists the results, where the best are shown in bold, and the second best are shown with an underline. We can observe that the gating

Table 4
ECE (%) and NLL results on AD classification on ResNet-18, ResNet-50, DenseNet-121, and InceptionV3. The results are shown with mean and standard deviation across folds (Mean±Std).

Models	Metric	Baseline	TFKD	PSKD	FitNets	FRSKD	BYOT	CReg-KD
ResNet-18	ECE↓	10.92±1.14	11.14±2.36	10.62±0.64	12.87±3.95	10.94±2.55	9.90±1.86	**9.84±2.55**
	NLL↓	0.41±0.08	0.42±0.09	0.33±0.05	0.37±0.21	0.40±0.07	0.34±0.09	**0.29±0.04**
ResNet-50	ECE↓	9.02±1.42	**7.75±1.40**	8.48±2.11	11.35±2.41	8.92±2.53	9.4 ± 2.32	8.07±2.63
	NLL↓	0.34±0.05	0.34±0.04	**0.31±0.03**	0.35±0.13	0.35±0.04	0.51±0.13	**0.31±0.03**
DenseNet-121	ECE↓	8.51±3.07	8.51±0.83	8.81±1.33	11.16±4.02	9.17±2.09	8.14±1.30	**8.07±1.27**
	NLL↓	0.39±0.07	**0.29±0.04**	0.33±0.05	0.37±0.22	0.35±0.08	0.40±0.22	**0.29±0.06**
InceptionV3	ECE↓	8.45±4.48	8.38±0.79	8.13±2.06	8.80±1.29	9.29±2.44	8.73±1.50	**7.92±1.99**
	NLL↓	0.36±0.03	0.36±0.06	0.39±0.04	0.39±0.08	0.35±0.04	0.54±0.13	**0.33±0.10**

Table 5
Ablation studies on the effect of the gating mechanism, output distillation, and feature distillation. The components involved are listed √. Results of accuracy are shown with mean and standard deviation (Mean±Std).

Gating	Output distillation	Feature distillation	ResNet-18↑	ResNet-50↑	DenseNet121↑	InceptionV3↑
			87.05±5.00	88.45±1.58	86.01±2.22	90.21±3.11
	√		87.75±4.38	89.51±2.51	91.96±1.18	92.65±1.17
		√	88.23±3.84	89.44±2.62	90.50±2.96	90.72±1.99
	√	√	89.94±2.33	89.97±2.30	91.25±1.22	91.60±1.03
√		√	91.40±2.05	90.62±2.27	92.20±1.78	92.66±0.64
√	√		90.93±3.45	90.28±1.70	91.15±1.23	92.66±0.57
√	√	*	91.84±2.08	91.55±2.31	91.90±1.38	93.00±1.75
√	√	√	**92.35±2.10**	**92.02±2.48**	**92.36±2.27**	**94.05±1.20**

* indicates the raw feature distillation without refinement.

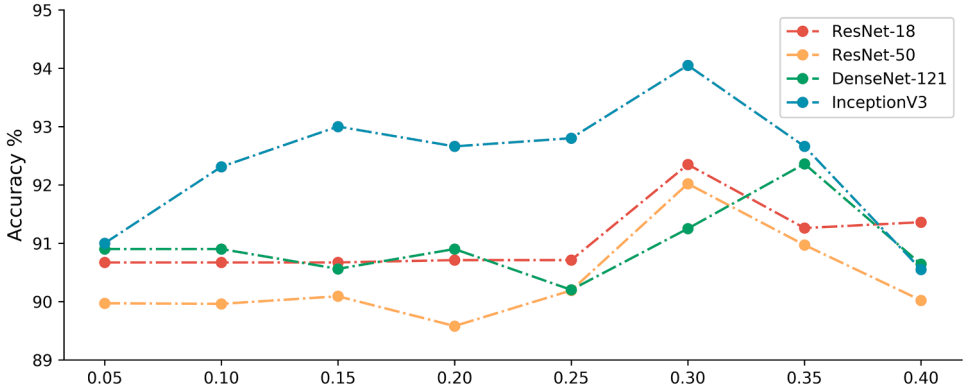

Fig. 3. The effect of the scale setting on ResNet-18 (red), ResNet-50 (orange), DenseNet-121 (green), and InceptionV3 (blue) models in terms of AD classification accuracy, where the scale ranged from 0.05 to 0.4. The best performances of the four architectures are achieved on the value of 0.3 or 0.35.

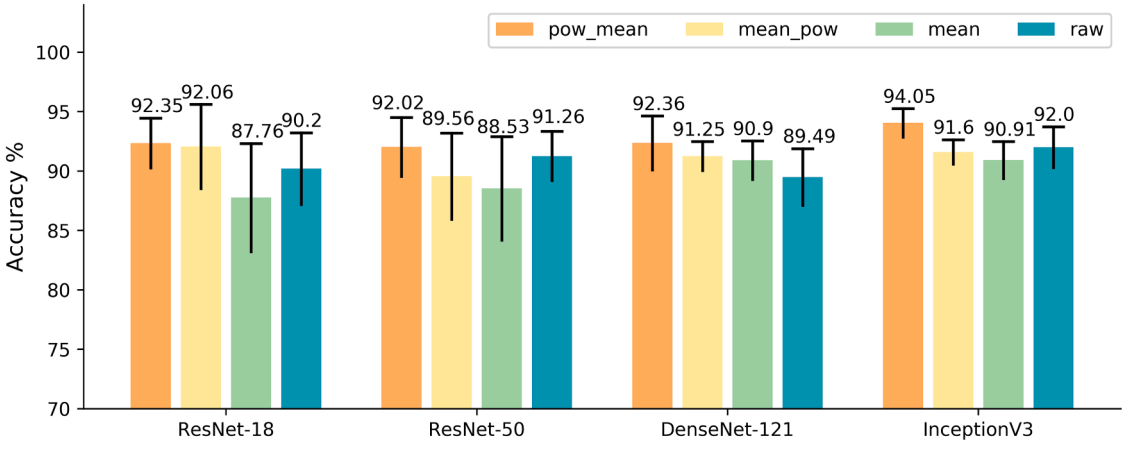

Fig. 4. The effect of the feature refinement on four architectures, including power and mean (orange), mean power (yellow), mean (light green), and raw features (dark green). The function ϕ_{pow_mean} outperforms other functions consistently.

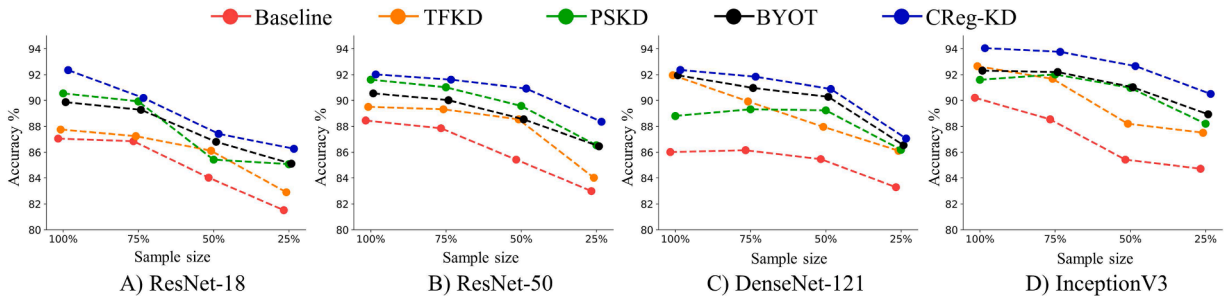

Fig. 5. Evaluation on the AD classification by training with different sizes of samples on A) ResNet-18, B) ResNet-50, C) DenseNet-121, and D) InceptionV3 by using various knowledge distillation regularization terms.

Table 6
Comparison results on brain age prediction in terms of mean absolute error (MAE), Pearson Correlation Coefficient (PCC), and Cumulative Score (CS%). The baseline column indicates the teacher model. The best results are shown in bold, and the second best are shown with an underline.

Models	Metric	Baseline	TFKD	PSKD	FitNets	FRSKD	BYOT	CReg-KD
ResNet-18	MAE↓	2.689±0.019	2.592±0.013	2.648±0.015	2.607±0.026	2.528±0.020*	2.466±0.032*	**2.413±0.013***
	PCC↑	0.982	0.984	0.985	0.985	0.985	0.985	**0.987***
	CS↑	85.352	85.070	85.915	85.915	88.169	**88.451**	88.169
ResNet-50	MAE↓	2.552±0.016	2.514±0.034	2.510±0.018	2.576±0.021	2.593±0.012	2.434±0.018*	**2.391±0.016***
	PCC↑	0.987	0.986	0.987	0.986	0.986	0.985	**0.988**
	CS↑	86.478	86.479	86.761	86.761	85.915	**88.887**	88.451
SFCN	MAE↓	2.342±0.020	2.345±0.015	2.325±0.039	2.315±0.025*	2.377±0.013	2.341±0.015	**2.217±0.026***
	PCC↑	0.988	0.989	0.989	0.990	0.989	0.989	**0.991**
	CS↑	85.955	87.921	87.921	87.640	87.640	89.326	**91.292**
InceptionV3	MAE↓	2.349±0.026	2.316±0.026*	2.364±0.021	2.345±0.016	2.385±0.026	2.295±0.027*	**2.162±0.027***
	PCC↑	0.986	0.988	0.987	0.987	0.985	0.988*	**0.989***
	CS↑	88.600	89.296	87.605	89.014	89.014	88.169	**90.986**

* : significantly outperforming with p-value < 0.05.

Table 8
KL divergence results on brain age classification on ResNet-18, ResNet-50, SFCN, and DeepBrainNet. The results denote the predicted errors, which are the differences between predicted and manually designed distribution. A lower value indicates better confidence estimation.

Models	Metric	Baseline	TFKD	PSKD	FitNets	FRSKD	BYOT	CReg-KD
ResNet-18	KL↓	0.0140	0.0127	0.0126	0.0124	0.0126	0.0129	**0.0116**
ResNet-50	KL↓	0.0117	0.0125	0.0121	0.0130	0.0119	0.0125	**0.0115**
SFCN	KL↓	0.0116	0.0109	0.0111	**0.0102**	0.0119	0.0108	**0.0102**
DeepBrainNet	KL↓	0.0122	0.0114	0.0120	0.0113	0.0122	0.0115	**0.0108**

Table 7
Ablation studies on the effect of the gating mechanism, output distillation, and feature distillation for brain age estimation. The components involved are listed with √. Results are shown with mean and standard deviation.

Gating	Output distillation	Feature distillation	ResNet-18↓	ResNet-50↓	SFCN↓	DeepBrainNet↓
			2.689±0.019	2.552±0.016	2.342±0.020	2.349±0.026
		√	2.549±0.019	2.547±0.029	2.359±0.021	2.369±0.024
	√		2.592±0.013	2.514±0.034	2.345±0.015	2.316±0.026
	√	√	2.677±0.016	2.466±0.021	2.310±0.031	2.349±0.021
√		√	2.515±0.014	2.579±0.011	2.274±0.015	2.171±0.022
√	√		2.577±0.019	2.421±0.016	2.301±0.018	2.282±0.024
√	√	*	2.508±0.022	2.402±0.027	2.234±0.019	2.168±0.021
√	√	√	**2.413±0.013**	**2.391±0.016**	**2.217±0.026**	**2.162±0.027**

* indicates the raw feature distillation without refinement.

mechanism plays an important role in improving performance, which corresponds to the results in the AD classification task. In addition, taking the output and feature distillation together might even decrease the performance. For example, with output and feature distillation, ResNet-18 without gating performs worse than using output or feature distillation alone. Nevertheless, when we combine all the components, further improvements are achieved in all the architectures. Additionally, we conducted a comparison between CReg-KD with refined feature distillation (the second line from the bottom) and Gated-KD (the first line from the bottom) (Yang et al., 2021) with raw feature distillation, with their performances listed in the second row from the bottom. The results indicate that the inclusion of the attentive refinement layer does not have a detrimental effect on performance. Compared with raw feature distillation, slight improvements are achieved on four architectures by CReg-KD. In particular, Gated-KD and CReg-KD achieves comparable results on the DeepBrainNet model. This suggests that the models may have reached their performance limit for the brain age prediction.

.

4.3.4. Evaluation on different sizes of training samples

We also assessed the performance of the four architectures in predicting brain age using different sample sizes in Fig. 6. Consistent with the results in Section 4.2.4, the four architectures trained with our proposed CReg-KD consistently achieved the best performance among all the knowledge distillation methods. In comparison, other knowledge distillation approaches, i.e. TFKD, PSKD, and BYOT, outperformed the baseline but achieved similar performance across all architectures. Furthermore, our CReg-KD demonstrated significant improvements in brain age prediction performance. The consistent improvement in both AD classification and brain age prediction tasks highlights the generalization ability and robustness of our proposed CReg-KD across various tasks. Overall, our findings underscore the effectiveness of CReg-KD in improving performance across different architectures for brain age prediction. Additionally, the consistent performance improvements in both AD classification and brain age prediction tasks provide further evidence of the versatility and reliability of our proposed CReg-KD approach.

5. Discussion

In this study, we proposed a Confidence Regularized Knowledge Distillation (CReg-KD) for filtering teacher knowledge and refining intermediate representations with attentive semantic contexts. The results demonstrate that the CReg-KD can efficiently improve the performance as well as generalization ability by carrying out two experiments: AD classification and brain age estimation. Extensive evaluations reveal that CReg-KD outperforms most existing state-of-the-art knowledge distillation approaches, and facilitates providing more accurate confidence estimation. Specifically, InceptionV3 improved the average accuracy from 90.21% to 93.05% in AD classification, and DeepBrainNet achieved an MAE of 2.162 in brain age estimation. The results demonstrate the superiority of the proposed CReg-KD as a powerful regularization paradigm for brain imaging analysis.

On the other hand, apart from TFKD and FitNet, the CReg-KD framework regularizes the student model learning by penalizing both the softmax output and intermediate representations. From Table 1 and Table 4, we can see that distillation on either the softmax output or hint features could achieve improvements on different models. Nevertheless, PSKD performed better than FitNets on ResNet-18 (90.55% vs. 87.39%) and ResNet-50 (91.61% vs. 88.81%), while worse on DenseNet-121 (88.80% vs. 89.50%) and InceptionV3 (91.60% vs. 92.65%) for AD classification. The same thing also happened in the brain age prediction task. From this point of view, it is still unclear to sense whether to perform distillation on the softmax output or the intermediate representations. One way to address this issue is by penalizing them together. However, by assessing the ablation study results in Table 5 and Table 7, we can see that the performance of distillation on both softmax output and hint features (the fourth row) might be worse than distillation on only one (the second or the third row). For example, on the AD classification task, the InceptionV3 model with both output and feature distillation achieves 91.60% accuracy on average, which is lower than that using output distillation alone with an accuracy of 92.65%. We suspect that this may be resulted from over-regularization due to an inappropriate selection of intermediate representations. And the increased dimensions of input might raise the computational burden for mimicking high-dimensional intermediate feature representations. In addition, once the gating mechanism is onset, the issue would be mitigated. One main reason is that the gating of knowledge is an attentive function that adaptively mediates the regularization paradigm in student learning, and softens the transferred knowledge to provide a trade-off for knowledge transfer.

In this study, we carried out experiments on four architectures for two tasks, including the baseline model ResNet18, and the state-of-the-art backbone InceptionV3. Interestingly, in AD classification, the simple model ResNet-18 with CReg-KD performs better than the teacher model by using InceptionV3. Moreover, it even outperforms the Dual Attention Multi-Instance Deep Learning model (AD vs. NC: 92.4%) (Zhu et al., 2021), which is a state-of-the-art model typically designed for AD classification. Although the data preprocessing pipeline and enrolled participants are not entirely the same, the results demonstrate the power of self-supervised knowledge distillation for improving performance in medical image classification to some extent. Overall, the proposed CReg-KD framework is powerful to improve model performance and generalization ability for medical image analysis.

This study has some limitations. Our proposed CReg-KD is implemented based on a pre-trained teacher model and transfers knowledge from the teacher to the student model. Previous studies have demonstrated that this teacher model can be leveraged by past predictions at a certain epoch (Kim et al., 2021) or the student model itself (Zhang et al., 2019). One of our future works is to leverage these paradigms with our proposed CReg-KD framework for improving training efficiency, as well as remaining comparable performances. Moreover, the gating scale value plays a key role in the performance and is decided by a grid search. In the future, we would investigate an efficient way of measuring the gating scale value.

6. Conclusion

Learning with limited samples as well as achieving promising performance and generalization is a challenging task in medical image analysis. In this study, we revisit the knowledge distillation technology as a regularization paradigm by introducing additional knowledge to reinforce student learning. Accordingly, we propose a confidence-regularized knowledge distillation framework and demonstrate its feasibility and generalization ability on two tasks: AD classification and brain age estimation. By investigating the confidence of transferred knowledge and the semantics of representations, our proposed CReg-KD

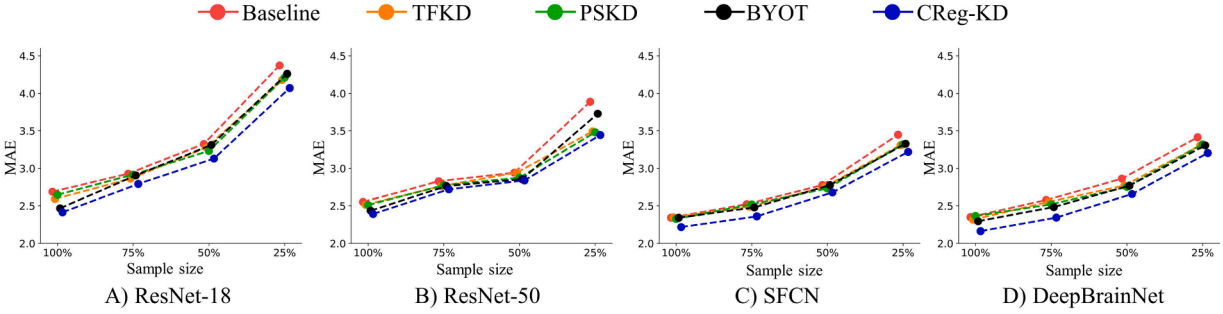

Fig. 6. Evaluation on the brain age prediction by training with different sizes of samples on A) ResNet-18, B) ResNet-50, C) SFCN, and D) DeepBrainNet by using various knowledge distillation regularization terms.

provides a way of attentive regularization for penalizing distilled knowledge in both intermediate representations and softmax output. Extensive experimental results show the superiority of CReg-KD in achieving consistent improvements over baseline models and outperforming other state-of-the-art knowledge distillation methods.

CRediT authorship contribution statement

Yanwu Yang: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Xutao Guo:** Investigation, Visualization, Validation. **Chenfei Ye:** Data curation, Formal analysis, Funding acquisition. **Yang Xiang:** Writing – review & editing, Supervision, Funding acquisition. **Ting Ma:** Methodology, Writing – review & editing, Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data availability

I have shared the link of the data used in the paper.

Acknowledgments

We acknowledge Mindsgo Life Science Shenzhen Co.Ltd. for technical support on image data management and processing on the BrainSite cloud platform (http://brainsite.cn/). This study is supported by grants from the National Natural Science Foundation of China (62106113, 62276081, and 62106115), Guangdong Basic and Applied Basic Research Foundation (2023A1515010792), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (NO: ZYYCXTD-C-202004), the National Key Research and Development Program of China (2021YFC2501202), and the Major Key Project of Peng Cheng Laboratory.

Data availability

The code for the CReg-KD pipeline is available at https://github.com/podismine/CReg-KD. The datasets are publicly available online:

IXI: http://brain-development.org
ADNI: https://ida.loni.usc.edu/
FCP-1000: http://www.nitrc.org/projects/fcon_1000
OASIS: https://www.oasis-brains.org/

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2023.102916.

References

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., 2016. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1313–1321.

Bashyam, V.M., Erus, G., Doshi, J., Habes, M., Nasrallah, I.M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., 2020. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. Brain 143 (7), 2312–2324.

Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E., 2017. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart, pp. 111–119.

Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E., 2019. Phiseg: capturing uncertainty in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 119–127.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Chandrasegaran, K., Tran, N.-T., Zhao, Y., Cheung, N.-M., 2022. Revisiting label smoothing and knowledge distillation compatibility: what was missing?. In: International Conference on Machine Learning, pp. 2890–2916.

Cho, J.H., Hariharan, B., 2019. On the efficacy of knowledge distillation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4793–4801. https://doi.org/10.1109/ICCV.2019.00489.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning.

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V.C., Shi, L., Heng, P.-A., 2016. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Trans. Med. Imaging 35 (5), 1182–1195.

Dou, Q., Liu, Q., Heng, P.A., Glocker, B., 2020a. Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imaging 39 (7), 2415–2425. https://doi.org/10.1109/TMI.2019.2963882.

Dou, Q., Liu, Q., Heng, P.A., Glocker, B., 2020b. Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imaging 39 (7), 2415–2425.

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., Geng, X., 2017. Deep label distribution learning with label ambiguity. IEEE Trans. Image Process. 26 (6), 2825–2838.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: achievements and challenges. J. Digit. Imaging 32 (4), 582–596.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Huang, G., Liu, S., Van der Maaten, L., Weinberger, K.Q., 2018. Condensenet: an efficient densenet using learned group convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2752–2761.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18 (2), 203–211.

Islam, M., Glocker, B., 2021. Spatially varying label smoothing: capturing uncertainty from expert annotations. In: International Conference on Information Processing in Medical Imaging, pp. 677–688.

Ji, M., Shin, S., Hwang, S., Park, G., Moon, I.-C., 2021. Refine myself by teaching myself: feature refinement via self-knowledge distillation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10659–10668. https://doi.org/10.1109/CVPR46437.2021.01052.

Jónsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., Ellingsen, L.M., Walters, G.B., Gudbjartsson, D.F., Stefansson, H., Stefansson, K., Ulfarsson, M.O., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nat. Commun. 10 (1), 5409.

Kim, K., Ji, B., Yoon, D., Hwang, S., 2021. Self-knowledge distillation with progressive refinement of targets. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6547–6556. https://doi.org/10.1109/ICCV48922.2021.00650.

Liao, L., Zhang, X., Zhao, F., Lou, J., Wang, L., Xu, X., Zhang, H., Li, G., 2020. Multi-branch deformable convolutional neural network with label distribution learning for fetal brain age prediction. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 424–427.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., 2017. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. 35, 250–269.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J. Cogn. Neurosci. 22 (12), 2677–2684.

Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Vilaplana Besler, V., 2022. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation: analysis of ranking scores and benchmarking results. J. Mach. Learn. Biomed. imaging 1–54.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Mercan, C., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2017. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans. Med. Imaging 37 (1), 316–325.

Müller, R., Kornblith, S., Hinton, G., 2019. When does label smoothing help? Proceedings of the 33rd International Conference on Neural Information Processing Systems 4694–4703.

Nandakumar, N., Hsu, D., Ahmed, R., Venkataraman, A., 2022. DeepEZ: a Graph convolutional network for automated epileptogenic zone localization from resting-state fMRI connectivity. IEEE Trans. Biomed. Eng.

Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. 68, 101871.

Peters, M.E., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. ArXiv Preprint ArXiv: 1705.00108.

Qi, L., Kuen, J., Gu, J., Lin, Z., Wang, Y., Chen, Y., Li, Y., Jia, J., 2021. Multi-scale aligned distillation for low-resolution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14443–14453.

Rahimpour, M., Bertels, J., Radwan, A., Vandermeulen, H., Sunaert, S., Vandermeulen, D., Maes, F., Goffin, K., Koole, M., 2021. Cross-modal distillation to improve MRI-based brain tumor segmentation with missing MRI sequences. IEEE Trans. Biomed. Eng.

Ran, C., Yang, Y., Ye, C., Lv, H., Ma, T., 2022. Brain age vector: a measure of brain aging with enhanced neurodegenerative disorder specificity. Hum. Brain Mapp. 43 (16), 5017–5031.

Razzak, M.I., Naz, S., Zaib, A., 2018. Deep learning for medical image processing: overview, challenges and the future. Classification in BioApps 323–350.

Reinhold, J.C., Dewey, B.E., Carass, A., Prince, J.L., 2019. Evaluating the impact of intensity normalization on MR image synthesis. Med. Imaging 2019: Image Process. 10949, 890–898.

Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., & Bengio, Y. (2015). *FitNets: hints for thin deep nets* (arXiv:1412.6550). arXiv. http://arxiv.org/abs/1412.6550 [cs].

Shen, Y., Xu, L., Yang, Y., Li, Y., Guo, Y., 2022. Self-distillation from the last mini-batch for consistency regularization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11933–11942. https://doi.org/10.1109/CVPR52688.2022.01164.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Song, J., Chen, Y., Ye, J., Song, M., 2022. Spot-adaptive knowledge distillation. IEEE Trans. Image Process. 31, 3359–3370.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Venkataraman, A., Rathi, Y., Kubicki, M., Westin, C.-F., Golland, P., 2011. Joint modeling of anatomical and functional connectivity for population studies. IEEE Trans. Med. Imaging 31 (2), 164–182.

Wang, J., Zhang, F., Jia, X., Wang, X., Zhang, H., Ying, S., Wang, Q., Shi, J., Shen, D., 2022. Multi-class ASD classification via label distribution learning with class-shared and class-specific decomposition. Med. Image Anal. 75, 102294.

Wong, K.C., Syeda-Mahmood, T., Moradi, M., 2018. Building medical image classifiers with very limited data using segmentation networks. Med. Image Anal. 49, 105–116.

Yang, Y., Xutao, G., Ye, C., Xiang, Y., & Ma, T. (2021). Regularizing brain age prediction via gated knowledge distillation. *Med. Imaging Deep Learn.*.

Ye, Y., Zhang, J., Chen, Z., Xia, Y., 2022. DeSD: self-supervised learning with deep self-distillation for 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 545–555.

Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J., 2020. Revisiting knowledge distillation via label smoothing regularization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3902–3910. https://doi.org/10.1109/CVPR42600.2020.00396.

Yun, S., Park, J., Lee, K., Shin, J., 2020. Regularizing class-wise predictions via self-knowledge distillation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13873–13882. https://doi.org/10.1109/CVPR42600.2020.01389.

Zhang, L., Shi, Y., Shi, Z., Ma, K., Bao, C., 2020. Task-oriented feature distillation. Adv. Neural Inf. Process Syst. 33, 14759–14771.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K., 2019. Be your own teacher: improve the performance of convolutional neural networks via self distillation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3712–3721. https://doi.org/10.1109/ICCV.2019.00381Citation. Key: zhangBeYourOwn2019.

Zhang, M.-L., Zhou, Z.-H., 2013. A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. 26 (8), 1819–1837.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., & Zhang, Q. (2021). *Rethinking soft labels for knowledge distillation: a bias-variance tradeoff perspective* (arXiv: 2102.00650). arXiv. http://arxiv.org/abs/2102.00650 [cs].

Zhu, W., Sun, L., Huang, J., Han, L., Zhang, D., 2021. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. IEEE Trans. Med. Imaging 40 (9), 2354–2366.