Medical Image Analysis 84 (2023) 102725



Contents lists available at ScienceDirect

Medical Image Analysis

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





Adversarial learning-based multi-level dense-transmission knowledge distillation for AP-ROP detection

Hai Xie [a,1], Yaling Liu [b,1], Haijun Lei [c], Tiancheng Song [d], Guanghui Yue [a], Yueshanyi Du [b], Tianfu Wang [a], Guoming Zhang [b,*], Baiying Lei [a,*]

[a] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
[b] Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute, Shenzhen, China
[c] Guangdong Province Key Laboratory of Popular High-performance Computers, School of Computer and Software Engineering, Shenzhen University, Shenzhen, China
[d] Shenzhen Silan Zhichuang Technology Co., Ltd., Shenzhen, China

ARTICLE INFO

Keywords:
AP-ROP detection
Dense-transmission
Knowledge distillation
Adversarial learning

ABSTRACT

The Aggressive Posterior Retinopathy of Prematurity (AP-ROP) is the major cause of blindness for premature infants. The automatic diagnosis method has become an important tool for detecting AP-ROP. However, most existing automatic diagnosis methods were with heavy complexity, which hinders the development of the detecting devices. Hence, a small network (student network) with a high imitation ability is exactly needed, which can mimic a large network (teacher network) with promising diagnostic performance. Also, if the student network is too small due to the increasing gap between teacher and student networks, the diagnostic performance will drop. To tackle the above issues, we propose a novel adversarial learning-based multi-level dense knowledge distillation method for detecting AP-ROP. Specifically, the pre-trained teacher network is utilized to train multiple intermediate-size networks (i.e., teacher-assistant networks) and one student network by dense transmission mode, where the knowledge from all upper-level networks is transmitted to the current lower-level network. To ensure that two adjacent networks can distill the abundant knowledge, the adversarial learning module is leveraged to enforce the lower-level network to generate the features that are similar to those of the upper-level network. Extensive experiments demonstrate that our proposed method can realize the effective knowledge distillation from the teacher to student networks. We achieve a promising knowledge distillation performance for our private dataset and a public dataset, which can provide a new insight for devising light-weight detecting systems of fundus diseases for practical use.

1. Introduction

The Retinopathy of prematurity (ROP) usually exists in premature infants, which is the main cause of blindness (Brown et al., 2018). The characteristic of ROP is a vessel proliferative disorder of developing retinal vessels, which progresses quickly and causes irreversible visual impairment if not being intervened and treated in time. As described in (Campbell et al., 2020; Hutchinson, 2018), the World Health Organization in 2012 reported that 15 million babies are born premature annually, and a considerable number of these premature infants will have some physiological dysfunction, including the blindness caused by ROP. With the mild or moderate ROP without intervention, Aggressive Posterior Retinopathy of Prematurity (AP-ROP) will occur, which is one of the most serious types of ROP (Ahn et al., 2017). According to (Sen et al., 2020), screening is the primary step for the management of ROP. However, different regions or countries have different screening criteria and different ophthalmologists may have a subject difference on the judgment of ROP. Accurate identification is greatly beneficial for clinicians to make the corresponding scientific judgment and the choice of diagnosis and treatment methods. In particular, the incidence of AP-ROP is relatively low and many ophthalmologists have insufficient experience to distinguish the ROP and AP-ROP. An illustration of ROP, AP-ROP and normal fundus images is shown in Fig. 1. From Fig. 1, we can observe that the appearance and characteristic between ROP and

* Corresponding authors.
E-mail addresses: 13823509060@163.com (G. Zhang), leiby@szu.edu.cn (B. Lei).
[1] Indicates the joint first authors.

https://doi.org/10.1016/j.media.2022.102725
Received 21 November 2021; Received in revised form 31 October 2022; Accepted 2 December 2022
Available online 9 December 2022
1361-8415/© 2022 Elsevier B.V. All rights reserved.

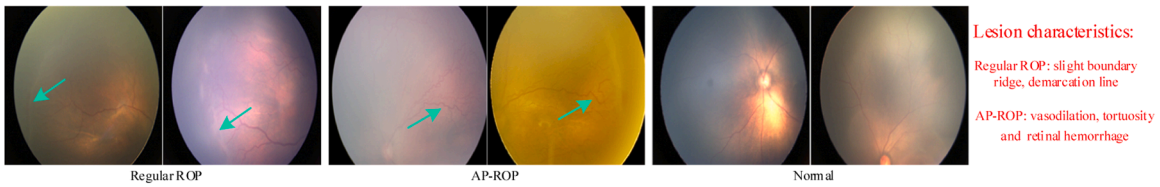


Fig. 1. An illustration of regular ROP, AP-ROP and Normal fundus images. The regular ROP images are expressed by slight boundary ridge and demarcation line. The appearance of AP-ROP is vasodilation, tortuosity and retinal hemorrhage.

AP-ROP is quite similar. The characteristics of some regular ROP are not obvious from the collected images, which are a big obstacle for ophthalmologists to identify them accurately and quickly. Hence, the computer-aided diagnosis becomes particularly important with respect to assisting the clinicians to make objective assessment of ROP and AP-ROP, which also attracts lots of attention (Campbell et al., 2020b, Redd et al., 2019). However, the existing methods are inefficient due to the high complexity, which cannot provide ophthalmologists a rapid feedback. Clinically, the quick screening is quite importance since it can help doctors make a rapid decision on the problematic cases from a large number of images to reduce the workload.

In recent years, deep learning technology has been developed rapidly, which leads to the continuous upgrade of computer hardware due to the high requirement from deep learning models. In practical applications, the deep learning models are usually required to have a small size with few parameters due to small storage and high-speed processing. Hence, some small models were developed to meet the deployment of small devices or mobile devices (such as series of MobileNet (Howard et al., 2019;Howard et al., 2017; Sandler et al., 2018), shuffleNet (Ma et al., 2018; Zhang et al., 2018)). However, these small models are insufficient to achieve good performance for all tasks due to the limited representation of deep high-level features. In order to solve this issue, Hinton et al. proposed the concept of knowledge distillation (KD) (Hinton et al., 2015), in which the whole deep learning network consists of two networks (i.e., teacher network with large size and student network with small size) and the student network is trained to learn the feature representation of teacher network by the constrained conditions. Although the existing KD methods have achieved great success in transmitting knowledge from teacher network to student network (Ju et al., 2021; Wang and Yoon, 2021), the large teacher networks do not always achieve good results.

According to (Mirzadeh et al., 2020), KD methods do not always perform well when the gap between the size of teacher network and student network is large (e.g., with the large size or capacity of teacher network, the small student network has a limited ability to mimic the feature representation from teacher network.). To solve this problem, the Teacher Assistant Knowledge Distillation (TAKD) framework is first proposed in (Mirzadeh et al., 2020), which introduces intermediate models as teacher assistant (TA) network to transmit the knowledge of teacher network to student network and fill the gap between the teacher and student network, where the knowledge of teacher model is distilled to TA models and the student model only mimics the TA models. Nevertheless, the obtained error from the distillation between upper-level and lower-level models will deliver directly all the time, which can cause the problem of error avalanche and further the student cannot learn the adequate information from the teacher model. To improve this shortcoming, inspired from DenseNet (Huang et al., 2017) that achieved excellent performance on image classification, Son et al. devised a densely guided KD model (Son et al., 2021), which employs multiple TA models and makes full use of the learned knowledge at each level so that the student can learn all front models instead of only the parent-level network. Despite the gained good classification performance by (Son et al., 2021), the presented model still misses some important feature information between a paired of the teacher-student network.

To solve the above issues and challenges, we propose a novel KD model, which contains a pair of teacher-student network and multiple TA networks. Similar to (Son et al., 2021), the outputs of these subnetworks are integrated via dense connection mode. The difference is that our model not only utilizes the output layer of all subnetworks, but also the last feature extraction layers of all subnetworks are leveraged to align and distillate knowledge from the upper-level network to the lower-level network. In particular, the adversarial learning mechanism is adopted to enforce the lower-level network to generate the features that are close to the features of the upper-level network, thereby realizing the feature alignment and compensate the information loss in the processing of KD from the upper-level network to the lower-level network. Extensive experiments demonstrate that our proposed framework can effectively guide the student network to learn knowledge from teacher and TA models for our private ROP dataset and a public dataset, which can ensure the small network to mimic the large network well and provide a new insight on the design of computer-aided diagnosis systems for detecting fundus diseases using color fundus images.

Overall, our main contributions are summarized as follows.

1) We devise a novel KD model based on multi-level dense-transmission TAKD networks and adversarial learning mechanism for detecting ROP and AP-ROP using color fundus images.
2) Both features of the last feature extraction layers and output layers from upper-level and lower-level networks are densely utilized in the processing of information transmission.
3) The adversarial learning mechanism is adopted to enforce the lower network to generate the features that are close to that of the upper-level network. To our best knowledge, this is the first framework that jointly utilizes the adversarial learning and dense TAKD models.

2. Related work

2.1. Automatic screening of ROP

For premature infants, ROP is a key factor of blindness (Hellström et al., 2013). If ROP can be found and treated in the early stage, it will greatly reduce the burden of family and society. For ophthalmologists, different clinicians may exist subjective differences for the same case and there is a general lack of sufficient experience of ophthalmologists around the world. Hence, many researchers attempted to study the automatic screening system for ROP, and some outstanding models were proposed (Campbell et al., 2021; Gupta et al., 2019; Redd et al., 2019). For example, Redd et al. presented a deep learning-based system to detect plus disease in ROP and a quantitative assessment of vessel vascular abnormality on a 1–9 scale is generated (Redd et al., 2019). Brown *et al.* used two convolution neural networks (CNNs) to complete the automatic diagnosis of plus disease in ROP, where one CNN employs U-Net architecture to realize the segmentation of retinal vessel and the other CNN adopted Inception-v1 structure to predict the category of the input images (Brown et al., 2018). Taylor et al. exploited a deep algorithm to monitor the disease progression with a quantitative severity scale for ROP (Taylor et al., 2019). However, these above models are difficult to deploy to a small device due to the high complexity.

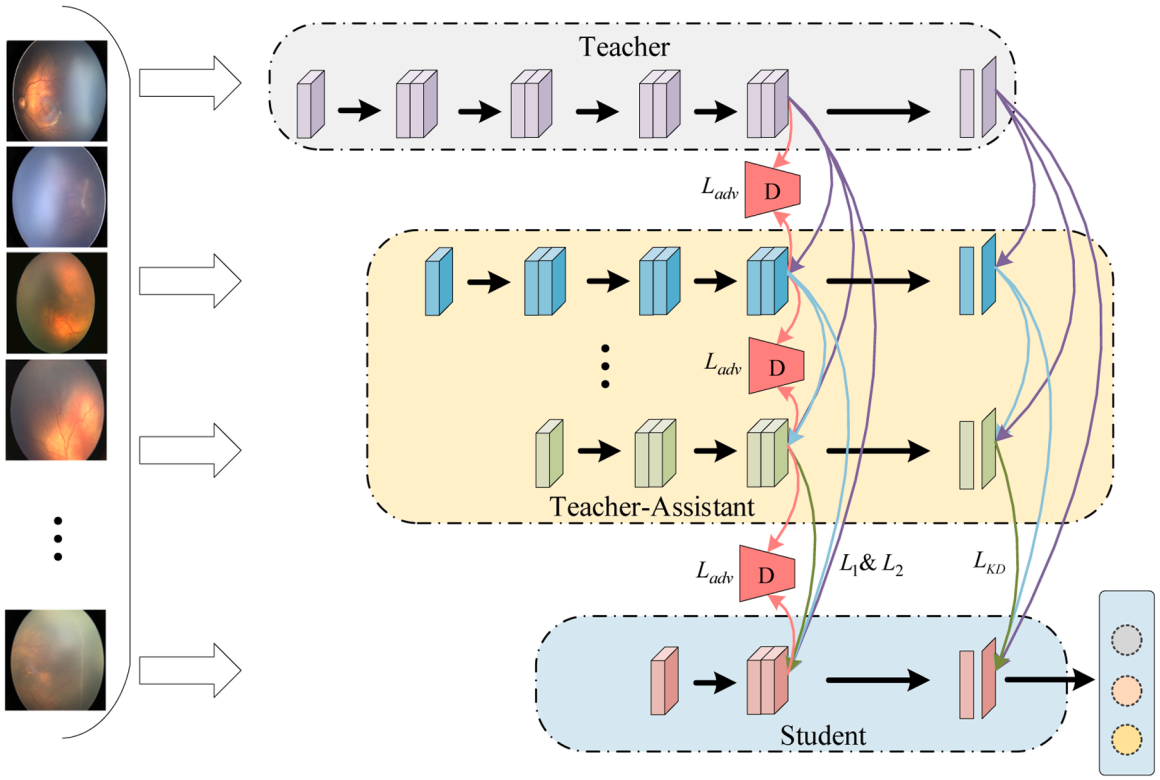


Fig. 2. Architecture of the proposed method. The whole framework consists of a paired of teacher-student network and multiple teacher-assistant networks. The knowledge of teacher with large size is distillated step by step to a small network and multiple teacher-assistant networks are employed to learn the knowledge from teacher network to reduce the big gap between the teacher and student network. To avoid heavy information loss, the output layer of the current network and that of all the front networks are computed by KD loss, and the feature of the last feature extraction layer are restrained by L1-norm and L2-norm. Particularly, in order to further reduce the feature information loss, the adversarial learning module is adopted to align the feature space between the two adjacent networks.

2.2. Knowledge distillation

With the popularity of mobile and portable devices, many deep learning-based applications tend to be developed with small size and low requirement for devices. This demand leads to the research upsurge of model compression (Buciluǎ et al., 2006; Deng et al., 2020), in which KD is a main topic (Li et al., 2020b; Wang et al., 2019a; Zhao et al., 2020; Zhao et al., 2021). The principle of KD is that the knowledge from teacher network with a large size is distilled to a student network with a small size to enable the student network to mimic the teacher network so that the small network also can realize a stronger feature expression capacity than itself. Since the explanation of KD was elaborated by (Hinton et al., 2015), plenty of KD models were proposed (Park et al., 2019; Wang et al., 2018, 2019b; Yao and Sun, 2020). For instance, Park et al. proposed a relational KD model, which transfers the relations of the outputs structure-wise to individual outputs themselves (Park et al., 2019). To strengthen the ability of distillation, Yao et al. presented a dense cross-layer mutual distillation method to promote knowledge representation learning and improve the capacity of transfer knowledge (Yao and Sun, 2020). Li et al. proposed a semi-supervised domain adaption method to accomplish the segmentation of annotation-efficient cardiac data, in which the student network can learn labeled target data, unlabeled target and labeled source data from multiple modalities simultaneously by the devised dual-teacher model (Li et al., 2020a). In addition, proved by (Mirzadeh et al., 2020), the capacity of student network will be reduced if the gap of network size between teacher and student network. For this reason, teacher-assistant networks (Mirzadeh et al., 2020; Son et al., 2021) are used as intermediate-size model to reduce the gap between the teacher and student networks.

2.3. Adversarial learning for KD

Since generative adversarial networks (GANs) were proposed by (Goodfellow et al., 2014), various GANs were devised to solve diverse tasks, such as image segmentation (Hu et al., 2020; Wang et al., 2019c), synthesis (He et al., 2021; Yang et al., 2019), classification (Tan et al., 2021; Xie et al., 2020) and domain adaption (Li et al., 2020c; Mahmood et al., 2018). In this paper, adversarial learning is adopted to enforce the lower-level networks to generate the features that are similar to those of the upper-level network. When a KD model transfers knowledge from teacher to student network, the loss of feature information is inevitable. Hence, some researchers explored the use of adversarial learning mechanism to align the feature spaces of teacher and student networks (Kundu et al., 2019; Liu et al., 2020; Nguyen-Meidine et al., 2021; Wang et al., 2018), for the reason that the adversarial learning strategy can enforce the student network to generate the features that are close to that of teacher network. For example, Shen et al. proposed an ensemble model to distill the knowledge from multiple pre-trained teacher networks, where the adversarial learning strategy is utilized to align feature spaces and guide the student network to recover the knowledge in teacher networks (Shen et al., 2019). Zhang et al. utilized adversarial learning mechanism to generate extra divergent examples to enhance the diversity of the transferred knowledge, and further enrich the predictions of multiple classifiers for the devised co-distillation phase (Zhang et al., 2021).

3. Methodology

To realize the accurate and quick screening of computer-aided diagnosis systems for ROP and AP-ROP, this paper proposes a novel KD model based on adversarial learning and multi-level dense-

Algorithm 1
Multi-level dense-transmission via KD model.

Input: ROP data x, ground-truth label y_{tr}, pre-trained teacher network T, pre-trained teacher-assistant network $TA_j, j = 1, 2, \ldots$, number of teacher-assistant network n, student network S, max epoch *epoch_max*.
Output: the distilled model S
  Define two empty lists:L_{feas} and L_{logits}
If $n \geq 1$ **then**
  for $i = 1$ **to** *epoch_max* **do**
    feed x into T to obtain the T_{fea}, T_o;
    append T_{fea}, T_o to L_{feas} , L_{logits};
    for $j = 1$ **to** n **do**
      feed x into TA_j to obtain the $TA_{j_{fea}}$, TA_{j_o};
      append $TA_{j_{fea}}$, TA_{j_o} to L_{feas} , L_{logits};
    end
    feed x into S to obtain S_{fea}, S_o;
    for $m = 1$ to n;
      Pop L_{feas} , L_{logits} from lists L_{feas} and L_{logits};
    end
    Update S by the equation (5);
  end
end

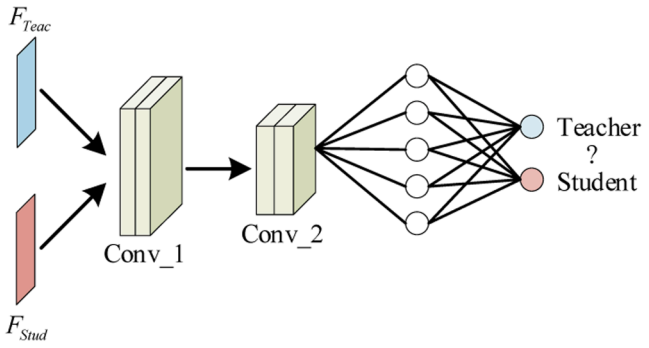


Fig. 3. Structure of our discriminator. F_{Teac} and F_{stud} represent the features of the last feature extraction layer from the teacher and student networks, respectively.

transmission. The whole framework of our proposed method is shown in Fig. 2. We will describe each module in detail in the following sections.

3.1. Principle of KD

The principle of KD is that the teacher network with a large size achieves a good performance using the trained dataset, while the student network with a small size is trained to mimic the capacity of the trained teacher network to obtain promising performance efficiently due to the design of fewer parameters. As the first prevalent description in (Hinton et al., 2015), the fully-connected layer output and logits (i.e., the softmax layer output) are used as the learned knowledge of teacher and student networks. Assume that T_o and S_o represent the output logits of teacher and student networks, respectively. y_{tr} denotes the ground-truth label. The KL divergence function is selected as the KD loss, which is defined as L_{KD}. The output logits of T_o and S_o are computed to obtain KD loss, which can be expressed as:

$$L_{KD} = \mathscr{T}^2 KL\left(softmax\left(\frac{T_0}{\mathscr{T}}\right), softmax\left(\frac{S_o}{\mathscr{T}}\right)\right), \tag{1}$$

where $\mathscr{T}$ is the temperature hyper-parameter to control the softened signals produced from the output of teacher and student networks. Except for KD loss, the output of student network and the ground-truth label y_{tr} is also computed by cross-entropy loss to supervise the generated information, which is defined as:

$$L_{CE} = \mathscr{H}(softmax(S_o), y_{tr}), \tag{2}$$

Overall, the final loss function is integrated by a balancing parameter λ, which can be written as follows:

$$L = (1 - \lambda)L_{CE} + \lambda L_{KD}, \tag{3}$$

3.2. Multi-level dense-transmission KD

As proved in (Mirzadeh et al., 2020), the big gap between the teacher and student network can lead to the student network learn insufficient knowledge from teacher network. If we want to design a small enough network to meet the requirement of clinical efficiency, there are a big

Table 1
The detailed information of our designed networks based on residual structure in experiments.

Layer name	Output size	18-layer	14-layer	10-layer	8-layer	6-layer
Conv1	112 × 112	7 × 7, 64, stride=2				
Conv2_x	56 × 56	3 × 3 Max-pooling, stride=2				
		$\begin{pmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{pmatrix} \times 3$	$\begin{pmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{pmatrix} \times 2$	$\begin{pmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{pmatrix} \times 1$	(3 × 3, 64) × 2	(3 × 3, 64) × 1
Conv3_x	28 × 28	$\begin{pmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{pmatrix} \times 3$	$\begin{pmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{pmatrix} \times 2$	$\begin{pmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{pmatrix} \times 1$	(3 × 3, 128) × 2	(3 × 3, 128) × 1
Conv4_x	14 × 14	$\begin{pmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{pmatrix} \times 1$	$\begin{pmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{pmatrix} \times 1$	$\begin{pmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{pmatrix} \times 1$	(3 × 3, 256) × 1	(3 × 3, 256) × 1
Conv5_x	7 × 7	$\begin{pmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{pmatrix} \times 1$	$\begin{pmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{pmatrix} \times 1$	$\begin{pmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{pmatrix} \times 1$	(3 × 3, 512) × 1	(3 × 3, 512) × 1
Pooling	1 × 1	Average-pooling, 7 × 7				
Output	2 or 3	Fully-connected, Softmax, 2 for task 1, 3 for task 2				

Table 2
The classification performance of different naïve baseline. #Params represents the number of the parameters of the corresponding models, M denotes million and the values in brackets indicate variance. (%).

Model	#Params	Task 1				Task 2			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
18-layer	11.2M	79.42 (0.04)	79.35 (0.04)	79.31 (0.03)	79.34 (0.03)	72.65 (0.03)	72.27 (0.26)	69.74 (0.18)	71.02 (0.05)
14-layer	5.3M	76.79 (0.09)	76.75 (0.07)	76.57 (0.11)	76.63 (0.09)	72.17 (0.11)	71.96 (0.19)	69.32 (0.16)	70.69 (0.13)
10-layer	4.9M	78.75 (0.14)	79.04 (0.21)	78.35 (0.13)	78.68 (0.14)	72.47 (0.01)	73.60 (0.23)	68.98 (0.22)	71.14 (0.04)
8-layer	1.9M	75.92 (0.02)	75.93 (0.01)	75.64 (0.05)	75.79 (0.02)	70.01 (0.07)	68.90 (0.03)	66.28 (0.24)	67.60 (0.09)
6-layer	1.7M	76.05 (0.31)	76.50 (0.08)	75.57 (0.55)	76.00 (0.22)	69.86 (0.08)	70.37 (0.28)	66.16 (0.23)	68.13 (0.07)

Table 3
The classification performance of different distillation methods without considering TA-KD models. #Params represents the number of the parameters of the corresponding models, M denotes million and black arrow → represents the direct distillation (%).

Method	#Params	Task 1				Task 2			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Teacher (ResNet-34)	21.3M	98.62 (0.004)	98.66 (0.003)	98.58 (0.005)	98.62 (0.004)	96.15 (0.01)	95.45 (0.01)	95.22 (0.03)	95.30 (0.01)
T_{34}→S_{18}	11.2M	97.91 (0.02)	97.94 (0.01)	97.86 (0.02)	97.91 (0.02)	95.88 (0.02)	95.21 (0.02)	94.99 (0.09)	95.10 (0.04)
T_{34}→S_{14}	5.3M	97.21 (0.06)	97.32 (0.08)	97.11 (0.06)	97.21 (0.07)	94.91 (0.03)	94.31 (0.02)	94.21 (0.14)	94.24 (0.03)
T_{34}→S_{10}	4.9M	96.43 (0.04)	96.51 (0.03)	96.34 (0.04)	96.37 (0.03)	93.46 (0.02)	92.61 (0.09)	92.01 (0.08)	92.25 (0.03)
T_{34}→S_{8}	1.9M	96.30 (0.16)	96.39 (0.11)	96.19 (0.20)	96.20 (0.09)	92.62 (0.03)	91.92 (0.03)	91.73 (0.11)	91.86 (0.03)
T_{34}→S_{6}	1.7M	94.24 (0.07)	94.28 (0.06)	94.16 (0.07)	94.15 (0.05)	92.34 (0.57)	91.43 (0.87)	91.04 (3.39)	91.21 (0.67)

Table 4
The detection performance of different distillation paths with TA-KD models. Red arrow → represents the distillation with TA-KD models (%).

Method	Task 1				Task 2			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Teacher (ResNet-34)	98.62 (0.004)	98.66 (0.003)	98.58 (0.005)	98.62 (0.004)	96.15 (0.01)	95.45 (0.01)	95.22 (0.03)	95.30 (0.01)
T_{34}	97.91 (0.02)	97.94 (0.01)	97.86 (0.02)	97.91 (0.02)	95.88 (0.02)	95.21 (0.02)	94.99 (0.09)	95.10 (0.04)
→TA_{18}								
T_{34}	97.23 (0.01)	97.30 (0.02)	97.15 (0.01)	97.21 (0.01)	95.72 (0.01)	95.01 (0.06)	95.04 (0.01)	95.02 (0.01)
→TA_{18}								
→TA_{14}								
T_{34}	96.83 (0.03)	96.97 (0.02)	96.71 (0.04)	96.83 (0.03)	95.23 (0.01)	94.67 (0.01)	94.43 (0.02)	94.55 (0.01)
→TA_{18}								
→TA_{14}								
→TA_{10}								
T_{34}	96.46 (0.03)	96.60 (0.02)	96.34 (0.03)	96.45 (0.02)	94.70 (0.05)	94.40 (0.37)	93.42 (0.30)	93.79 (0.02)
→TA_{18}								
→TA_{14}								
→TA_{10}								
→TA_{8}								
T_{34}	94.67 (0.19)	94.86 (0.22)	94.57 (0.20)	94.64 (0.17)	93.60 (0.01)	91.73 (0.05)	93.41 (0.10)	92.55 (0.02)
→TA_{18}								
→TA_{14}								
→TA_{10}								
→TA_{8}								
→S_{6}								

gap between the network and the teacher network, which will cause the learning ability to be greatly reduced. Hence, the prediction performance is undesired. To solve this issue, multiple teacher-assistant networks are utilized in (Mirzadeh et al., 2020; Son et al., 2021) as the intermediated-size networks to transfer the knowledge from teacher network to student network. Inspired by this idea, we propose a multi-level dense-transmission KD model. In this model, the ResNet-34 model being pre-trained on ImageNet dataset is chosen as the teacher network to train our ROP data and obtain the trained ResNet-34 model, which can be distilled to the teacher-assistant and student networks. As shown in Fig. 2, the current network can distill knowledge from all the front networks via the dense transmission mode. The mathematical definition can be described as follows.

We define the teacher-assistant networks as $TA_i,\ i = 1, 2, \ldots$ between the teacher network T and student network S, and the losses are denoted as L_{TA_i}. Thus, the loss of student network can be expressed as:

$$L_S = L_{T\to S} + L_{TA_i\to S}, \tag{4}$$

where the symbol → represents the distillation flow from the upper-level network. As shown in Fig. 2, the output features of the last feature layer of each network are also utilized to constrain the teacher-assistant and student networks by L1-norm and L2-norm. The features of the upper-level and lower-level network are indicated as f_{up} and f_{low}, respectively. So the loss of feature constraint can be defined as $L_1 = |f_{up} - f_{low}|^1$ and $L_2 = \|f_{up} - f_{low}\|^2$. After integrating the losses according to Eq. (3), the total loss for student network can be summarized as below.

$$\begin{aligned} L_s = {} & (1-\lambda)\sum_{i=1}^{n}\left(L_{CE_{T\to S}} + L_{CE_{TA_i\to S}}\right) \\ & + \lambda\sum_{i=1}^{n}\left(L_{KD_{T\to S}} + L_{KD_{TA_i\to S}} + L_{1_{T\to S}} + L_{1_{TA_i\to S}} + L_{2_{T\to S}} + L_{2_{TA_i\to S}}\right), \end{aligned} \tag{5}$$

Table 5
The detection performance of different distillation paths with multi-level dense-transmission KD. Green arrow → represents the multi-level dense-transmission KD (%).

Method	Task 1				Task 2			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Teacher (ResNet-34)	98.62 (0.004)	98.66 (0.003)	98.58 (0.005)	98.62 (0.004)	96.15 (0.01)	95.45 (0.01)	95.22 (0.03)	95.30 (0.01)
T_{34}	97.91 (0.02)	97.94 (0.01)	97.86 (0.02)	97.91 (0.02)	95.88 (0.02)	95.21 (0.02)	94.99 (0.09)	95.10 (0.04)
→TA_{18}								
T_{34}	97.54 (0.03)	97.62 (0.02)	97.46 (0.04)	97.57 (0.02)	95.82 (0.03)	95.23 (0.07)	94.83 (0.06)	95.09 (0.02)
→TA_{18}								
→TA_{14}								
T_{34}	97.02 (0.01)	97.16 (0.01)	96.90 (0.02)	97.02 (0.02)	95.75 (0.01)	95.21 (0.02)	94.62 (0.01)	94.91 (0.01)
→TA_{18}								
→TA_{14}								
→TA_{10}								
T_{34}	96.65 (0.01)	96.72 (0.01)	96.57 (0.01)	96.64 (0.01)	95.48 (0.04)	94.67 (0.08)	94.80 (0.03)	94.68 (0.03)
→TA_{18}								
→TA_{14}								
→TA_{10}								
→TA_{8}								
T_{34}	95.69 (0.01)	95.79 (0.01)	95.58 (0.01)	95.70 (0.01)	94.70 (0.01)	93.37 (0.08)	94.12 (0.03)	93.77 (0.01)
→TA_{18}								
→TA_{14}								
→TA_{10}								
→TA_{8}								
→S_{6}								

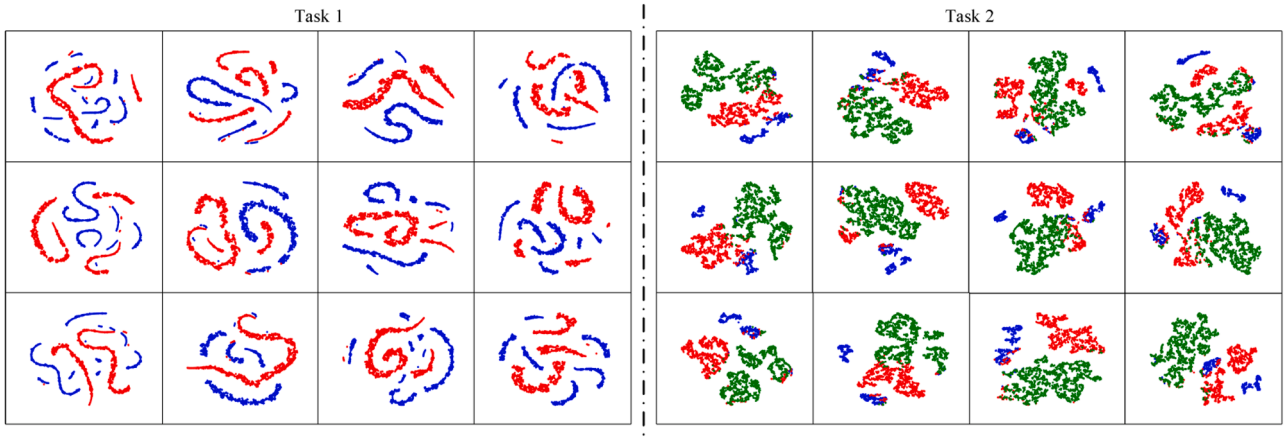


Fig. 4. The t-SNE visualization results of different distillation paths of the three distillation strategies for task 1 and task 2. From top to down: three distillation strategies without TA models, TA models and the proposed method. From left to right: distillation paths of 14-layer, 10-layer, 8-layer and 6-layer residual modules being chosen as student network.

Finally, the multi-level dense-transmission based KD flow can be summarized in Algorithm 1.

3.3. Adversarial learning for KD

In this paper, the adversarial learning mechanism is applied to the features of the last feature extraction layer from each network, which enforces the teacher-assistant or student network to generate the features that are close to the teacher network so that the feature information loss can be further reduced in the processing of distillation. As shown in Fig. 2, the discriminator is employed to distinguish the generated features whether are obtained from the upper-level network between two adjacent networks.

The used discriminator employs two convolution layers with the kernel size being set as 1, an adaptive average pooling layer, a fully-connected layer and a sigmoid layer. The structure of our proposed

Table 6
The comparative results between without and with adversarial learning for task 1 (%).

	Distillation paths	Acc	Pre	Rec	F1
No Discriminator	T_{34}	97.45 (0.04)	97.56 (0.03)	97.35 (0.05)	97.48 (0.04)
	$\rightarrow TA_{18}$				
	TA_{18}	97.29 (0.04)	97.39 (0.03)	97.20 (0.05)	97.25 (0.03)
	$\rightarrow TA_{14}$				
	TA_{14}	96.35 (0.02)	96.47 (0.02)	96.23 (0.02)	96.31 (0.01)
	$\rightarrow TA_{10}$				
	TA_{10}	95.07 (0.03)	95.14 (0.01)	94.97 (0.05)	95.06 (0.03)
	$\rightarrow TA_{8}$				
	TA_{8}	93.47 (0.01)	93.49 (0.01)	93.36 (0.01)	93.42 (0.01)
	$\rightarrow S_{6}$				
Discriminator	T_{34}	97.91 (0.02)	97.94 (0.01)	97.86 (0.02)	97.91 (0.02)
	$\rightarrow TA_{18}$				
	TA_{18}	97.54 (0.03)	97.62 (0.02)	97.46 (0.04)	97.57 (0.02)
	$\rightarrow TA_{14}$				
	TA_{14}	97.02 (0.01)	97.16 (0.01)	96.90 (0.02)	97.02 (0.02)
	$\rightarrow TA_{10}$				
	TA_{10}	96.11 (0.10)	96.26 (0.06)	95.99 (0.13)	96.10 (0.10)
	$\rightarrow TA_{8}$				
	TA_{8}	95.69 (0.01)	95.79 (0.01)	95.58 (0.01)	95.70 (0.01)
	$\rightarrow S_{6}$				

Table 7
The comparative results between without and with adversarial learning for task 2 (%).

	Distillation paths	Acc	Pre	Rec	F1
No Discriminator	T_{34}	95.52 (0.01)	94.98 (0.08)	94.84 (0.03)	94.89 (0.04)
	$\rightarrow TA_{18}$				
	TA_{18}	95.44 (0.01)	94.39 (0.04)	94.69 (0.02)	94.52 (0.01)
	$\rightarrow TA_{14}$				
	TA_{14}	95.32 (0.02)	94.73 (0.06)	94.52 (0.06)	94.66 (0.05)
	$\rightarrow TA_{10}$				
	TA_{10}	92.97 (0.01)	92.08 (0.12)	91.76 (0.03)	91.95 (0.02)
	$\rightarrow TA_{8}$				
	TA_{8}	88.99 (0.01)	86.87 (0.04)	87.58 (0.10)	87.26 (0.01)
	$\rightarrow S_{6}$				
Discriminator	T_{34}	95.88 (0.02)	95.21 (0.02)	94.99 (0.09)	95.10 (0.04)
	$\rightarrow TA_{18}$				
	TA_{18}	95.82 (0.03)	95.23 (0.07)	94.83 (0.06)	95.09 (0.02)
	$\rightarrow TA_{14}$				
	TA_{14}	95.75 (0.01)	95.21 (0.02)	94.62 (0.01)	94.91 (0.01)
	$\rightarrow TA_{10}$				
	TA_{10}	95.48 (0.04)	94.67 (0.08)	94.80 (0.03)	94.68 (0.03)
	$\rightarrow TA_{8}$				
	TA_{8}	94.70 (0.01)	93.37 (0.08)	94.12 (0.03)	93.77 (0.01)
	$\rightarrow S_{6}$				

discriminator is shown in Fig. 3. We define the j-level discriminator as D_j. The two adjacent networks are denoted as Net_j and Net_{j+1}, respectively, where $Net_j$is considered as the teacher network and Net_{j+1} indicate the student network to be trained. Thus, the adversarial loss function can be expressed as:

$$L_{adv} = \mathbb{E}_{x\sim p(Net_j)} log D_j(x) + \mathbb{E}_{x\sim p(Net_{j+1})} log\big(1 - D_j(x)\big). \quad (6)$$

Note that our discriminators are only adopted between two adjacent networks and the dense connection strategy is not applied to the discriminators.

4. Experiments

4.1. Materials and experimental setup

The used ROP dataset is captured by RetCam3 from a Shenzhen Eye hospital and the time span of the collection is from 2009 to 2018. This dataset consists of 13,508 color fundus images, which is labeled by three professional ophthalmologists. We abandon the images that the ophthalmologists hold different views and the images with poor quality. Eventually, we obtain 12,095 images for our experiments, which contain 1698 AP-ROP, 3898 regular ROP and 6499 normal fundus images. In the experiments, we utilize the 5-fold cross validation to obtain the results. In addition, we select the trained parameters of the model whose results are closest to the mean value as teacher or teacher-assistant network. Particularly, we utilize this dataset for two tasks, one is for detecting the ROP and normal fundus images (i.e., binary classification task) and the other one is for classifying AP-ROP and ROP from normal fundus images (i.e., three-class classification task). All the fundus images are resized to 224×224 as the input of all used models.

The Pytorch library is employed to train and test all the used models. One NVIDIA TITAN XP GPU is leveraged to accelerate the speed of training and testing. The Adam with the betas $= (0.9, 0.997)$, eps $= 10^{-8}$ is chosen as the optimizer for student network. The discriminators employ the SGD as the optimizer with a momentum of 0.9 and weight decay of 5×10^{-5}. The initial learning rates of student and discriminator networks are set to 0.001, which are multiplied by 0.1 every 30 and 20 epochs, respectively. The max epoch is set to 120 and the size of the mini-batch is set to 32. We employ accuracy (Acc), Precision (Pre), Recall (Rec), and F1-score (F1) as the evaluation metrics.

4.2. Experimental results

In our experiments, there are two tasks to be accomplished: detection of ROP and normal images for task 1, classification of AP-ROP, regular ROP and normal images for task 2. The task 1 can assist the clinicians to make a preliminary screening quickly. To further judge the severity of ROP the patients suffer from, the task 2 can help the clinicians to make a decision on the category of ROP so that the most appropriate treatment can be implemented in time. We mainly employ the residual structure to build the small networks. According to the number of residual structures, we design the networks of 18, 14, 10, 8, and 6 layers, in which the smallest model with 6 layers is selected as the student network and other networks are considered as teacher-assistant models. The detailed structure information can be found in Table 1.

On the basis of the devised network architectures, we first implement the naïve baseline to show the ability of extracting features without knowledge distillation. The experimental results are reported in Table 2.

From Table 2, we can see that the obtained results by the devised naïve network are not good. For the reason that these networks do not use pre-training parameters of large-scale dataset, which causes that these networks learn insufficient features. In order to resolve this issue, the KD algorithm is a good solution, which can distillate the knowledge from the large network with the pre-trained parameters on ImageNet to small network. By this way, the classification performance can be

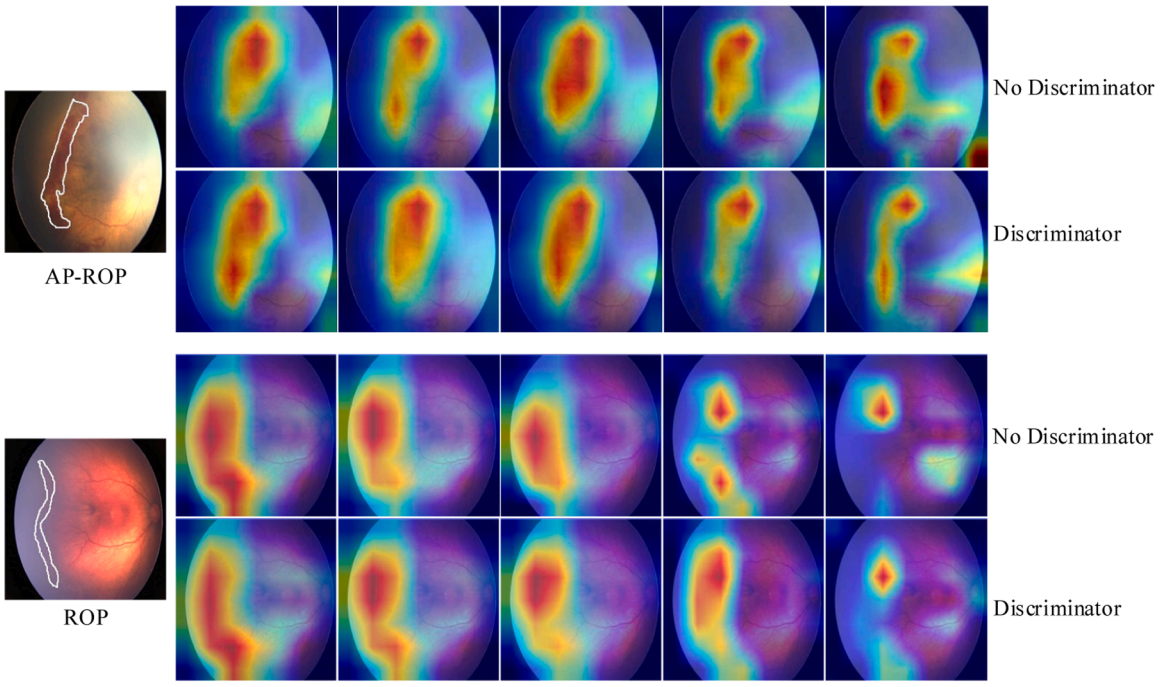


Fig. 5. The Grad-CAM visualization results of different distillation paths without and with the proposed discriminator. From left to right: distillation path of T_{34}→TA_{18}, TA_{18}→TA_{14}, TA_{14}→TA_{10}, TA_{10}→TA_8, TA_8→S_6. The first two rows represent the results of the comparison group for AP-ROP and the last two rows indicate the results of the comparison group for regular ROP, where the areas framed by white lines represent the lesion region.

boosted.

According to the experimental setting, the ResNet-34 model in (He et al., 2016) being pre-trained on ImageNet is selected as the teacher network and the TA_i (i is set as 18, 14, 10, 8 in our experiments) indicates the teacher-assistant models. However, in order to show the effectiveness of the devised model, we first use these teacher-assistant models as the student networks to conduct experiments, namely, S_i (i = 18, 14, 10, 8, 6) represent the student network that are utilized to distill knowledge from the teacher network directly. The detection performance of different distillation networks for task 1 is reported in Table 3, we can see that the obtained distillation results are quite good due to the devised adversarial learning module and loss function, which will be verified in the following ablation experiments. By comparing Tables 2 and 3, we can see that the classification performance has been greatly improved. For example, the accuracy of the task 1 and task 2 increase 18.19% and 22.48% for the 6-layer residual network, respectively. These improved results benefit by the teacher network (ResNet-34) using the pre-training parameters of ImageNet dataset and the usage of KD between teacher and student networks.

When the network size is shrunken by nearly 11 times, the accuracy is decreased by 4.38% and 3.81% for task 1 and task 2, respectively. In addition, as the size of the student network becomes smaller, the performance of the prediction becomes weaker, which verifies the view in (Mirzadeh et al., 2020) that the larger the gap between the teacher and student networks, the worse the obtained distillation performance. To illustrate the addition of TA models can compensate the reduction of performance caused by the gap between the teacher and student networks, we conduct a group of experiments using TA models as the intermediated-size models to distillate knowledge from teacher network to student network. The results are reported in Table 4, where TA_i (i = 18, 14, 10, 8) represent the corresponding size of the teacher-assistant network. The number of parameters is the same as that of Table 3, for the reason that only student network is trained, teacher and TA models provide the features of the last feature extraction layer and the output logits. From Table 4, we can see that all the evaluation metrics are improved with the design of the TA models between teacher and student networks, which proves the conclusion in (Mirzadeh et al., 2020) that the intermediate-size TA models can distillate knowledge from teacher to student network more efficient than only using teacher-student architecture. The classification can be further enhanced for both task 1 and task 2 due to the decreased gap with the use of TA models.

To demonstrate the effectiveness of our proposed Algorithm 1, we conduct a group of experiments with different teacher-assistant networks using the dense connection of the features from the last feature extraction layer and the output logits. The experimental results are reported in Table 5. We can see that all results are improved with the design of multi-level dense transmission KD when compared with the results obtained by only a single path (i.e., distillation knowledge from teacher network to student network directly with the use of TA models) that are shown in Table 4. For example, the accuracy of the path 'T_{34}→TA_{18}→TA_{14}→TA_{10}→TA_8 →S_6' is 1.0% higher than that of the path 'T_{34}→TA_{18}→TA_{14}→TA_{10}→TA_8 →S_6' for task 1 and there is also more than a percentage point increasing with respect to accuracy for task 2, which demonstrate that KD with TA models and adversarial learning modules by multi-dense transmission can reduce the loss of feature information and boost the overall classification performance.

Also, we conduct t-SNE visualization to compare the learned features by the used three distillation strategies, as shown in Fig. 4. The results are only from four paths, where the distillation paths of 14-layer, 10-layer, 8-layer and 6-layer residual modules are chosen as student networks. Since the above four paths can satisfy the condition of the multi-level dense-transmission KD mechanism. We can observe that compared with the first two rows, the last row can better separate different classes and cluster the same class.

4.3. Ablation studies

4.3.1. Adversarial learning

As the main contribution, the discriminator is an important module for boosting the performance of the baseline network, and further enhancing all the distillation paths. Hence, we conduct a comparative experiment without the integration of a discriminator to prove the

Table 8
The comparative results of ablation experiments on different loss function from different distillation strategies for task 1 and task 2. Black arrow → means the direct distillation paths without TA models, red arrow ⟶ indicates the distillation paths with TA models and green arrow ⟶ represents the multi-level dense-transmission KD (%).

L_{adv}	L_{KD}	L_1	L_2	Distillation Strategies	Task 1				Task 2			
					Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	√	√	√	T_{34}→S_6	92.41 (0.07)	92.75 (0.02)	92.17 (0.12)	92.42 (0.04)	86.92 (0.02)	85.27 (0.07)	85.03 (0.14)	85.22 (0.01)
				T_{34}	92.82 (0.11)	92.91 (0.08)	92.72 (0.15)	92.87 (0.05)	88.42 (0.05)	86.36 (0.12)	86.82 (0.10)	86.55 (0.03)
				⟶TA_{18}								
				⟶TA_{14}								
				⟶TA_{10}								
				⟶TA_8								
				⟶S_6								
				T_{34}	93.47 (0.01)	93.49 (0.01)	93.36 (0.01)	93.42 (0.01)	88.99 (0.01)	86.87 (0.04)	87.58 (0.10)	87.26 (0.01)
				⟶TA_{18}								
				⟶TA_{14}								
				⟶TA_{10}								
				⟶TA_8								
				⟶S_6								
√		√	√	T_{34}→S_6	92.74 (0.02)	92.90 (0.01)	92.57 (0.03)	92.73 (0.02)	89.44 (0.02)	88.08 (0.02)	88.75 (0.07)	88.44 (0.02)
				T_{34}	95.19 (0.05)	95.47 (0.03)	94.99 (0.06)	95.20 (0.02)	91.39 (0.01)	90.49 (0.01)	90.07 (0.04)	90.29 (0.01)
				⟶TA_{18}								
				⟶TA_{14}								
				⟶TA_{10}								
				⟶TA_8								
				⟶S_6								
				T_{34}	95.62 (0.08)	95.81 (0.03)	95.47 (0.11)	95.62 (0.05)	93.31 (0.01)	92.43 (0.11)	91.55 (0.15)	92.06 (0.02)
				⟶TA_{18}								
				⟶TA_{14}								
				⟶TA_{10}								
				⟶TA_8								
				⟶S_6								
√	√		√	T_{34}→S_6	93.11 (0.02)	93.20 (0.01)	92.98 (0.03)	93.10 (0.01)	87.72 (0.03)	86.85 (0.13)	85.10 (0.08)	85.98 (0.05)
				T_{34}	94.53 (0.33)	94.85 (0.16)	94.32 (0.42)	94.61 (0.23)	87.87 (0.01)	86.22 (0.01)	85.38 (0.02)	85.79 (0.01)
				⟶TA_{18}								
				⟶TA_{14}								
				⟶TA_{10}								
				⟶TA_8								
				⟶S_6								
				T_{34}	95.08 (0.08)	95.30 (0.04)	94.91 (0.10)	95.08 (0.05)	93.62 (0.03)	92.91 (0.09)	92.67 (0.05)	92.72 (0.02)
				⟶TA_{18}								
				⟶TA_{14}								
				⟶TA_{10}								
				⟶TA_8								

(continued on next page)

Table 8 (*continued*)

L_{adv}	L_{KD}	L_1	L_2	Distillation Strategies	Task 1				Task 2			
					Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
				→S_6								
√	√	√		T_{34}→S_6	81.55 (0.03)	81.53 (0.02)	81.41 (0.06)	81.43 (0.03)	75.13 (0.02)	74.30 (0.11)	71.29 (0.21)	72.75 (0.03)
				T_{34}	91.54 (0.08)	91.76 (0.14)	91.37 (0.09)	91.54 (0.08)	86.54 (0.03)	83.82 (0.35)	85.97 (0.97)	84.93 (0.15)
				→TA_{18}								
				→TA_{14}								
				→TA_{10}								
				→TA_8								
				→S_6								
				T_{34}	92.30 (0.03)	92.39 (0.04)	92.17 (0.02)	92.27 (0.03)	88.38 (0.03)	86.30 (0.10)	87.86 (0.05)	86.97 (0.02)
				→TA_{18}								
				→TA_{14}								
				→TA_{10}								
				→TA_8								
				→S_6								
√	√	√	√	T_{34}→S_6	94.24 (0.07)	94.28 (0.06)	94.16 (0.07)	94.15 (0.05)	92.34 (0.57)	91.43 (0.87)	91.04 (3.39)	91.21 (0.67)
				T_{34}	94.67 (0.19)	94.86 (0.22)	94.57 (0.20)	94.64 (0.17)	93.60 (0.01)	91.73 (0.05)	93.41 (0.10)	92.55 (0.02)
				→TA_{18}								
				→TA_{14}								
				→TA_{10}								
				→TA_8								
				→S_6								
				T_{34}	95.69 (0.01)	95.79 (0.01)	95.58 (0.01)	95.70 (0.01)	94.70 (0.01)	93.37 (0.08)	94.12 (0.03)	93.77 (0.01)
				→TA_{18}								
				→TA_{14}								
				→TA_{10}								
				→TA_8								
				→S_6								

effectiveness of the adversarial learning mechanism for our multi-level network. The results of task 1 and task 2 are shown in Tables 6 and 7, respectively. In addition, we only display the results of the two adjacent networks, for the reason that the adversarial learning module is only adopted for the two adjacent networks.

From Tables 6 and 7, we can see that the results of most distillation paths without adversarial learning module are more than 0.5% less than that of the corresponding distillation paths with adversarial learning module. For example, the Acc, Pre, Rec and F1 of distillation path TA_{14}→TA_{10} without the devised discriminator are 0.67%, 0.69%, 0.67%, 0.71% lower than that of distillation path TA_{14}→TA_{10} with the proposed discriminator for task 1, respectively. For the distillation path TA_{10}→TA_8 for task 2, the Acc, Pre, Rec and F1 are reduced by 2.51%, 2.59%, 3.04% and 2.73% due to the absence of the discriminator. In particular, for the student network with 6-layer residual modules, the performance degradation is even worse than other teacher-assistant networks with the deletion of the designed discriminator. For instance, the accuracy of the distillation path TA_8→S_6 without discriminator is decreased by 5.71% compared with the path with discriminator for task 2. These results demonstrate that the adversarial learning module can effectively align the learned feature of the shallow network to the adjacent deep network and reduce the information loss during the distillation from the teacher network to the student network, especially for the large gap between the teacher and student networks.

To illustrate the good feature learning ability of shallow networks with the addition of the discriminator, we conduct the Grad-CAM visualization with the class activation maps (CAM) with the learned features from the last feature extraction layer for shallow networks. The Grad-CAM visualization results are shown in Fig. 5. By observing the comparison results, we can discover that our proposed model with discriminator can better learn the disease-specific features than the model without discriminator. For example, for distillation path T_{34}→TA_{18}, the teacher-assistant network can learn more information from teacher network when compared with other distillation paths for AP-ROP and regular ROP. In particular, the network with discriminator can learn finer feature than the network without discriminator. However, with the reduction of network size, the feature imitation ability of small-scale networks is weakening, and the learned information related

Table 9
The results on validation set of DDR dataset using three different distillation strategies. Black arrow → means the direct distillation paths without TA models, red arrow → indicates the distillation paths with TA models and green arrow → represents the multi-level dense-transmission KD (%).

Distillation strategies	Distillation paths	OA	AA	Kappa	F1
Direct distillation	Teacher (ResNet-34)	84.08	66.03	75.43	63.81
	T_{34}→S_{18}	83.64	68.96	74.50	63.40
	T_{34}→S_{14}	82.36	63.94	72.52	61.43
	T_{34}→S_{10}	81.74	65.70	71.30	60.36
	T_{34}→S_{8}	79.66	65.08	67.80	59.06
	T_{34}→S_{6}	77.97	57.45	65.65	54.44
Distillation with TA models	Teacher (ResNet-34)	84.08	66.03	75.43	63.81
	T_{34} →TA_{18}	83.64	68.96	74.50	63.40
	T_{34} →TA_{18} →TA_{14}	82.91	67.45	73.23	61.76
	T_{34} →TA_{18} →TA_{14} →TA_{10}	82.29	67.46	72.28	62.33
	T_{34} →TA_{18} →TA_{14} →TA_{10} →TA_{8}	80.53	66.83	69.29	64.91
	T_{34} →TA_{18} →TA_{14} →TA_{10} →TA_{8} →S_{6}	78.56	62.65	66.79	59.22
Multi-level dense distillation with adversarial learning	Teacher (ResNet-34)	84.08	66.03	75.43	63.81
	T_{34} →TA_{18}	83.64	68.96	74.50	63.40
	T_{34} →TA_{18} →TA_{14}	83.35	65.16	74.07	62.15
	T_{34} →TA_{18} →TA_{14} →TA_{10}	82.84	66.58	73.24	62.28
	T_{34} →TA_{18} →TA_{14} →TA_{10} →TA_{8}	81.30	62.69	70.71	58.91
	T_{34} →TA_{18} →TA_{14} →TA_{10} →TA_{8} →S_{6}	80.28	65.29	69.49	61.10

to disease types is also decreasing. Despite this, compared with the network without discriminator, the network with discriminator can learn more useful information. For instance, TA_8→S_6 (the last column in Fig. 5), there are some of the lesion features being not learned by employing the proposed model without adversarial learning module for AP-ROP and regular ROP. This phenomenon is even worse for the selected regular ROP fundus image, for the reason that the lesion characteristics of this fundus image is fuzzy and the contrast is quite low so that the student network cannot learn adequate features and generate the accurate features as the extracted feature by teacher network without the proposed discriminator.

4.3.2. Loss function

As described in Section 3, multiple loss functions are employed to optimize the training processing. In order to illustrate the effectiveness of the used loss functions with respect to boosting the performance, we conduct several ablation experiments by using the student network with the 6-layer residual structure. The experimental results are reported in Table 8, where the symbol '√' indicates this loss function is adopted in the training stage and the definition of distillation strategies is the same as that of Tables 3, 4, and 5.

As shown in Table 8, compared with the proposed method in which all the loss functions are adopted in the training stage, all classification results of task 1 and task 2 are reduced when one of these loss functions are removed for the compared three distillation strategies, especially for the deletion of adversarial loss and MSE loss. For example, for distillation path T_{34}→S_6, the accuracy of the model without MSE loss function is reduced by more than 10% than the model with MSE loss function for task 1 and task 2. However, the difference of accuracy is reduced greatly with the devised multi-level dense-transmission KD for task 1 and task 2 with respect to whether the MSE loss function is employed or not (i.e., the difference accuracy of task1 and task 2 is 3.39% and 6.32%, respectively). In addition, we can observe that the performance of the TA models is better than that of the models without using TA models as intermediate-size models and the performance of the multi-level dense-transmission KD model is better than those models using TA models. It demonstrates that the gap between teacher and student networks exists. The proposed multi-level dense-transmission KD model can effectively alleviate the information loss in the process of KD caused by the large gap between teacher and student networks.

5. Discussion

5.1. Validation

In this section, we conduct a series of experiments on a public dataset to verify the effectiveness of the proposed model. The public dataset is collected from multiple medical centers, called DDR, which is used to implement the tasks of diabetic retinopathy (DR) grading, lesion segmentation and object detection. According to the description of (Li et al., 2019), the DDR contains 13,673 color fundus images from 9598 patients

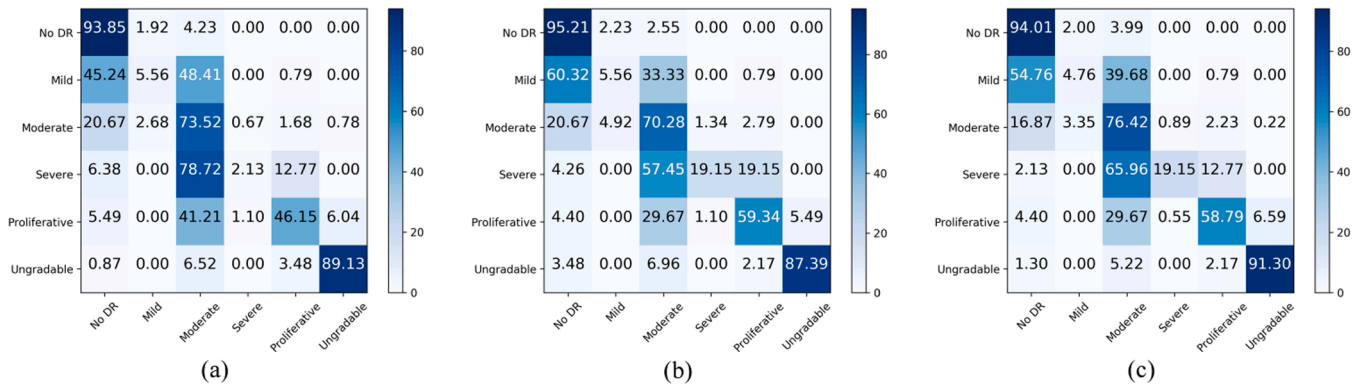


Fig. 6. The confusion matrices of the distillation paths of $T_{34}\rightarrow S_6$ (a), $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$ (b), $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$ (c).

by 147 medical centers. Also, the collected images are captured by 42 type of fundus cameras with 45°FOV. Hence, these conditions of the original images bring the great challenge for the DR recognition. In our experiments, we only conduct experiments on DDR dataset in terms of the task of DR grading to prove the effectiveness of the design of the proposed model. In (Li et al., 2019), the grading of DR is defined as six classes: no DR, mild DR, moderate DR, severe DR, proliferative DR, and ungradable. For the partition of DDR dataset, we keep the same settings as (Li et al., 2019), i.e., 50% for training, 20% for validation, 30% for testing. Thus, the training, validation and testing dataset contains 6835, 2733, 4105 images, respectively. For the experimental settings, all images are resized to 224 × 224, the batch size is set as 32. In the processing of training, the input images are augmented by color augmentation, random flipping and rotation. As described in (Li et al., 2019), the results on validation and testing are both reported for DR grading. Therefore, as the verification dataset, we choose the results on validation dataset to verify the effectiveness of the proposed distillation strategy. The experimental results are reported in Table 9. In (Li et al., 2019), the overall accuracy (OA), average accuracy (AA), Kappa are selected as the evaluation metrics. In addition to the above metrics, we add the F1-score (F1) to evaluate the proposed distillation strategies.

As reported in (Li et al., 2019), the presented model achieved 82.84%, 62.66% and 74.38% of OA, AA and Kappa, respectively. Nevertheless, we utilize the ResNet-34 model being pre-trained ImageNet dataset as the Teacher network to train the DDR dataset and obtain 84.08%, 66.03%, 75.43% and 63.81% of OA, AA, Kappa, and F1 respectively. We analyze the reason that the used data augmentation methods that consist of color augmentation, random flipping and random rotation can improve the classification performance. From Table 9, we can also see that the distillation strategy of direct distillation misses a large amount of information when the knowledge is distillated from T_{34} to S_6, in which the OA and Kappa decrease by 6.11% and 9.78%, respectively. By observing the distillation path $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$, the OA, AA, Kappa and F1 increase 0.59%, 5.2%, 1.14% and 4.78% with the distillation strategy of the addition of TA models, respectively. The results demonstrate that the TA models can boost the classification by reducing the loss of information, which also verifies the idea of (Mirzadeh et al., 2020). When the distillation strategy is switched to the multi-level dense distillation with adversarial learning, the classification performance is further improved. Through the distillation path of $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$, we can see that the metrics of OA, AA, Kappa and F1 increase 2.31%, 7.84%, 3.84% and 6.66%, respectively compared with the distillation path of $T_{34}\rightarrow S_6$, and the metrics of OA, AA, Kappa and F1 increase 1.72%, 2.64%, 2.7% and 1.88%, respectively compared with the distillation path of $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$.

To order to intuitively show the improvement of the proposed distillation strategy for different classes, we plot the confusion matrices with respect to the distillation path $T_{34}\rightarrow S_6$, $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$ and $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$, as shown in Fig. 6.

By observing the confusion matrices, the distillation strategy with TA models can greatly improve the accuracy of the classes of Severe and Proliferative so that the overall classification can be boosted. When using the strategy of multi-level dense distillation with adversarial learning, besides the classes of Severe and Proliferative, the classes of Moderate and Ungradable are also improved. Thereby, these results verify the effectiveness of the proposed distillation strategy.

5.2. The influence of batch-size

The different batch-size will affect the overall classification performance. Hence, we implement a pair of experiments to select the appropriate batch-size by setting different batch-sizes. The experimental results are reported in Table 10. In these comparative experiments, the results of the distillation paths of $T_{34}\rightarrow S_6$, $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$ and $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$ are utilized to prove the feasibility of the used batch-size in this paper. In Table 10, when the batch-size is set to 16 or 64, the overall performance is lower than that of the batch-size being set to 32. For example, for distillation path $T_{34}\rightarrow TA_{18}\rightarrow TA_{14}\rightarrow TA_{10}\rightarrow TA_8\rightarrow S_6$, the accuracy of the batch-size=32 is 1.41%, 0.05% higher than that of the batch-size=64 for task1 and task2, respectively. Hence, in our experiments, we set the batch-size as 32.

5.3. Limitation and future work

Although our devised teacher-student model can effectively transmit the knowledge of teacher network to student network and is better than the student network with a small size, which achieves a promising performance for detecting AP-ROP, there are some prediction failure examples, as shown in Fig. 7. From Fig. 7, we can observe that there exist some images with non-discriminatory characteristics, which leads to the false classification. For example, the images of the first two columns of the first row are considered as the regular ROP as the characteristics in these two images are similar to that of the regular ROP. Through the second row, we can see that the categories of AP-ROP and Normal are with the feature of regular ROP, which causes these images being erroneously classified to the type of regular ROP. In addition, we pick 100 images from the private ROP dataset that contains regular ROP, AP-ROP and Normal class, and visit three ophthalmologists to discriminate these images. As a result, the average accuracy of the three classes is

Table 10
The results of different batch-size. Black arrow → means the direct distillation paths without TA models, red arrow → indicates the distillation paths with TA models and green arrow → represents the multi-level dense-transmission KD (%).

Batch-size	Distillation Strategies	Task 1				Task 2			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
16	T_{34}→S_6	93.75 (0.03)	93.81 (0.03)	93.69 (0.04)	93.71 (0.02)	91.76 (0.04)	91.43 (0.11)	90.26 (0.88)	90.84 (0.11)
	T_{34}	94.21 (0.05)	94.27 (0.04)	94.11 (0.07)	94.23 (0.04)	93.53 (0.06)	93.45 (0.52)	91.93 (0.44)	92.84 (0.10)
	→TA_{18}								
	→TA_{14}								
	→TA_{10}								
	→TA_8								
	→S_6								
	T_{34}	95.24 (0.02)	95.52 (0.01)	95.05 (0.03)	95.29 (0.01)	94.64 (0.02)	94.06 (0.02)	93.38 (0.06)	93.70 (0.02)
	→TA_{18}								
	→TA_{14}								
	→TA_{10}								
	→TA_8								
	→S_6								
32	T_{34}→S_6	94.24 (0.07)	94.28 (0.06)	94.16 (0.07)	94.15 (0.05)	92.34 (0.57)	91.43 (0.87)	91.04 (3.39)	91.21 (0.67)
	T_{34}	94.67 (0.19)	94.86 (0.22)	94.57 (0.20)	94.64 (0.17)	93.60 (0.01)	91.73 (0.05)	93.41 (0.10)	92.55 (0.02)
	→TA_{18}								
	→TA_{14}								
	→TA_{10}								
	→TA_8								
	→S_6								
	T_{34}	95.69 (0.01)	95.79 (0.01)	95.58 (0.01)	95.70 (0.01)	94.70 (0.01)	93.37 (0.08)	94.12 (0.03)	93.77 (0.01)
	→TA_{18}								
	→TA_{14}								
	→TA_{10}								
	→TA_8								
	→S_6								
64	T_{34}→S_6	93.48 (0.04)	93.61 (0.03)	93.35 (0.06)	93.48 (0.04)	91.92 (0.01)	91.38 (0.26)	90.56 (0.18)	91.04 (0.01)
	T_{34}	94.15 (0.12)	94.48 (0.06)	93.93 (0.15)	94.24 (0.07)	93.47 (0.02)	92.10 (0.25)	92.82 (0.09)	92.37 (0.02)
	→TA_{18}								
	→TA_{14}								
	→TA_{10}								
	→TA_8								
	→S_6								
	T_{34}	94.28 (0.17)	94.60 (0.08)	94.05 (0.22)	94.33 (0.14)	94.65 (0.03)	94.06 (0.02)	93.45 (0.12)	93.73 (0.03)
	→TA_{18}								
	→TA_{14}								
	→TA_{10}								
	→TA_8								
	→S_6								

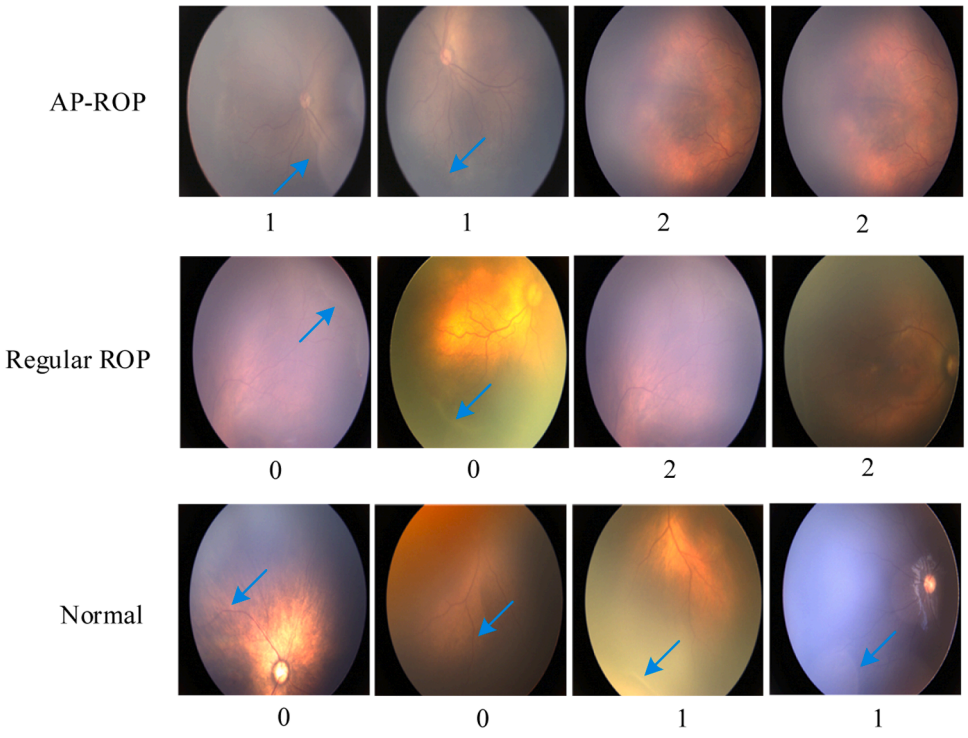


Fig. 7. Failure examples by the proposed distillation strategy. 0, 1, 2 represent the class of AP-ROP, regular ROP and Normal, respectively.

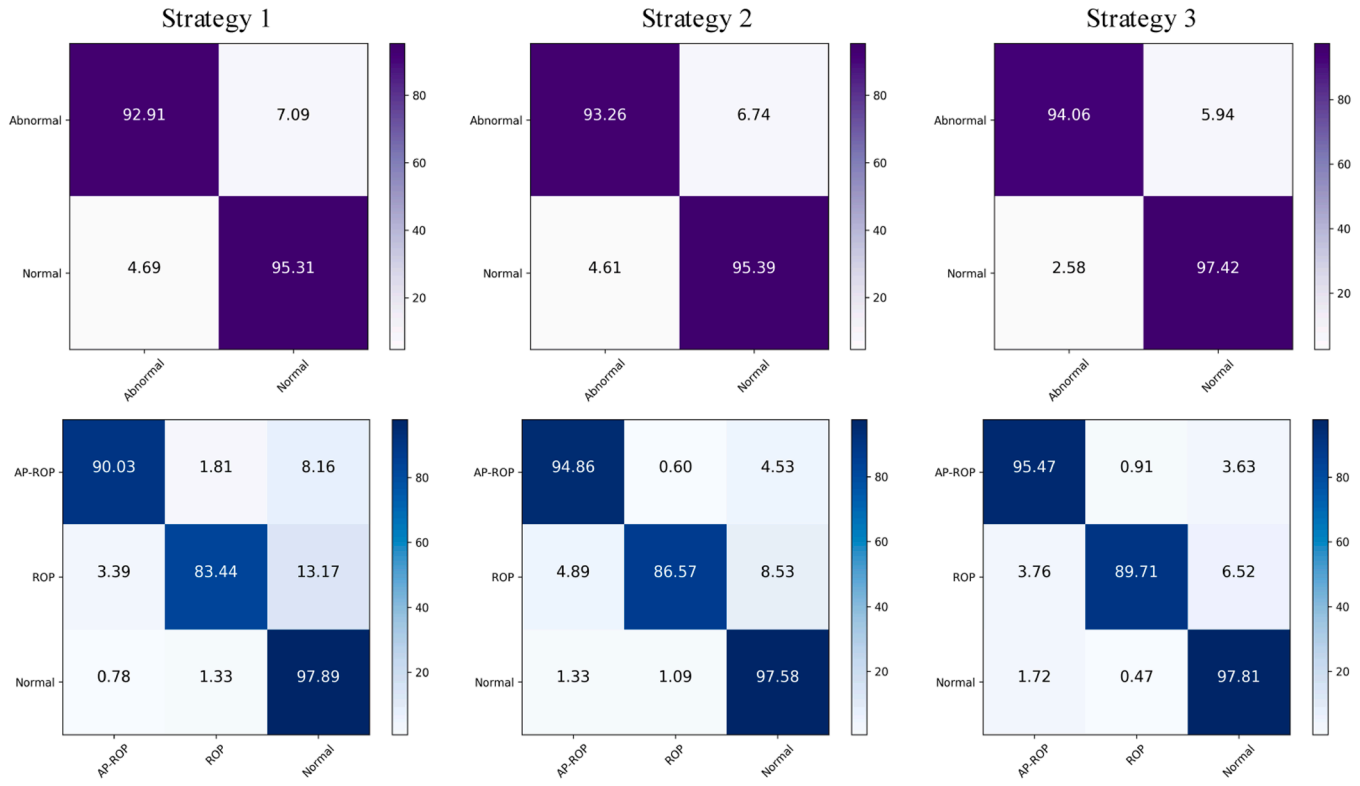


Fig. 8. Confusion matrixes of the used three distillation strategies in our experiments, where the 6-layer residual structure is selected as the student network. The upper and lower rows indicate the results of task 1 and task 2, respectively.

almost 85%, which is lower than 94.70% of the proposed method. The results demonstrate that our method has not only achieved satisfactory objective results, but also has a faster speed than doctors. However, there are still some limitations in this paper. For example, in order to extract discriminative features, we adopt multiple residual structures to build the TAKD models and student network, which causes the parameters still a little more than the plain network with a few layers. The teacher network only uses the ResNet-34 model to distillate knowledge

to shallow networks, which results in the limited performance of the teacher network. In addition, the data from just one center used to train and test the proposed method is limited. To address the above-mentioned limitations, we will make the following efforts in future work. First, some plain networks with smaller size should be devised to achieve more efficient distillation. Second, multiple teachers with excellent performance should be integrated to improve the detection performance of shallow networks. Thirdly, we will collect the data from multiple centers and explore the characteristics of different center data to enhance the generalization of our proposed method.

6. Conclusions

In this paper, we propose an adversarial learning-based multi-level dense-transmission KD model to avoid the information loss in the process of KD due to the large gap between teacher and student networks. Specifically, the ResNet-34 model being pre-trained on ImageNet dataset is utilized to train ROP data so that the trained teacher network can be obtained. The residual structures with different sizes are employed as the intermediate-size TA models and the network with a 6-layer residual structure is used as the student network to mimic the feature extraction capability of the teacher network. In particular, the multi-level dense-transmission KD mechanism is devised to reduce the feature loss in the process of KD by densely transmitting the learned feature information from the upper-level network to the low-level network so that the gap between the teacher and student networks can be compensated and the feature information loss can be alleviated. In addition, the adversarial learning module is applied to the adjacent networks to guide the shallow network to generate the features close to the features from the deep network. Hence, the feature information loss can be further reduced. Extensive experimental results demonstrate that the devised student network can mimic the teacher network well with the design of the multi-level dense-transmission KD and adversarial learning modules.

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

The authors do not have permission to share data.

Acknowledgements

This work was supported partly by National Natural Science Foundation of China (Nos. 62106153, U22A2024, 82271103 and 62001302), China Postdoctoral Science Foundation (2021M692196), Guangdong Basic and Applied Basic Research Foundation (Nos. 2020A1515110605, 2022A1515012326, 2021A1515011348 and 2020B121201001), Shenzhen Key Medical Discipline Construction Fund (No. SZXK038), Shenzhen Fund for Guangdong Provincial High level Clinical Key Specialties (No. SZGSP014), Shenzhen-Hong Kong Co-financing Project (No. SGDX20190920110403741), Shenzhen Key Basic Research Project (Nos. JCYJ20190808145011259 and JCYJ20220818095809021), Shenzhen Science and Technology Program (No. JSGG20201102174200001), Postdoctoral Innovation Achievement Funding Program of Health Science Center of Shenzhen University.

Appendix. Confusion Matrix

To show the superiority of the proposed model intuitively, we plot a comparative confusion matrix of the three distillation strategies in Tables 3, 4, and 5, where the 6-layer residual structure is chosen as the student network. In order to show the results clearly, we define the distillation strategy in Tables 3, 4, and 5 as Strategy 1, Strategy 2 and Strategy 3, respectively, as shown in Fig. 8. We can observe clearly that compared with the direct distillation models and TA models, the proposed method achieves obvious improvement for task 1 and task 2. The experimental results demonstrate that the multi-level dense-transmission KD mode can continuously use the knowledge from previous networks in the process of distillation, which can effectively avoid the information loss and ensure that the student can learn enough knowledge from the teacher network so that the classification performance can be boosted.

References

Ahn, Y., Hong, K., Yum, H., Lee, J., Kim, K., Youn, Y., Park, S., 2017. Characteristic clinical features associated with aggressive posterior retinopathy of prematurity. Eye 31 (6), 924–930.

Brown, J.M., Campbell, J.P., Beers, A., Chang, K., Ostmo, S., Chan, R.P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J., 2018. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA ophthalmol 136 (7), 803–810.

Buciluǎ, C., Caruana, R., Niculescu-Mizil, A., 2006. Model compression. In: Proc. SIGKDD, pp. 535–541.

Campbell, J.P., Chan, R.V.P., Ostmo, S., Anderson, J., Singh, P., Kalpathy-Cramer, J., Chiang, M.F., 2020. Analysis of the relationship between retinopathy of prematurity zone, stage, extent and a deep learning-based vascular severity scale. Invest. Ophthalmol. Vis. Sci 61 (7), 2193-2193.

Campbell, J.P., Kim, S.J., Brown, J.M., Ostmo, S., Chan, R.P., Kalpathy-Cramer, J., Chiang, M.F., Sonmez, K., Schelonka, R., Jonas, K., 2021. Evaluation of a Deep Learning–Derived Quantitative Retinopathy of Prematurity Severity Scale. Ophthalmol 128 (7), 1070–1076.

Deng, L., Li, G., Han, S., Shi, L., Xie, Y., 2020. Model compression and hardware acceleration for neural networks: a comprehensive survey. Proc. IEEE 108 (4), 485–532.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Gupta, K., Campbell, J.P., Taylor, S., Brown, J.M., Ostmo, S., Chan, R.P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J., 2019. A quantitative severity scale for retinopathy of prematurity using deep learning to monitor disease regression after treatment. JAMA ophthalmol 137 (9), 1029–1036.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. CVPR, pp. 770–778.

He, W., Sun, Y., Yang, M., Ji, F., Li, C., Xu, R., 2021. Multi-goal multi-agent learning for task-oriented dialogue with bidirectional teacher–student learning. Knowl-Based Syst 213, 106667.

Hellström, A., Smith, L.E., Dammann, O., 2013. Retinopathy of prematurity. The lancet 382 (9902), 1445–1457.

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

Howard, A., Sandler, M., Chu, G., Chen, L.-.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., 2019. Searching for mobilenetv3. Proc. CVPR 1314–1324.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, X., Guo, R., Chen, J., Li, H., Waldmannstetter, D., Zhao, Y., Li, B., Shi, K., Menze, B., 2020. Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images. J. Biomed. Health. Inform 24 (9), 2599–2608.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proc. CVPR, pp. 4700–4708.

Hutchinson, A.K., 2018. Retinopathy of prematurity revisited. JAMA ophthalmol 136 (12), 1389–1390.

Ju, L., Wang, X., Zhao, X., Lu, H., Mahapatra, D., Bonnington, P., Ge, Z., 2021. Synergic Adversarial Label Learning for Grading Retinal Diseases via Knowledge Distillation and Multi-task Learning. J. Biomed. Health. Inform. 25 (10), 3709–3720.

Kundu, J.N., Lakkakula, N., Babu, R.V., 2019. Um-adapt: unsupervised multi-task adaptation using adversarial cross-task distillation. In: Proc. CVPR, pp. 1436–1445.

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H., 2019. Diagnostic Assessment of Deep Learning Algorithms for Diabetic Retinopathy Screening. Inform. Sci. 501, 511–522.

Li, K., Yu, L., Wang, S., Heng, P.-A., 2020a. Towards cross-modality medical image segmentation with online mutual knowledge distillation. Proc. AAAI 34, 775–783.

Li, T., Li, J., Liu, Z., Zhang, C., 2020b. Few sample knowledge distillation for efficient network compression. In: Proc. CVPR, pp. 14639–14647.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020c. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med. Image Anal. 65, 101765.

Liu, Y., Shu, C., Wang, J., Shen, C., 2020. Structured knowledge distillation for dense prediction. IEEE Trans. Pattern. Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2020.3001940.

Ma, N., Zhang, X., Zheng, H.-.T., Sun, J., 2018. Shufflenet v2: practical guidelines for efficient cnn architecture design. In: Proc. ECCV, pp. 116–131.

Mahmood, F., Chen, R., Durr, N.J., 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imag. 37 (12), 2572–2581.
Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H., 2020. Improved knowledge distillation via teacher assistant. Proc. AAAI 34, 5191–5198.
Nguyen-Meidine, L.T., Belal, A., Kiran, M., Dolz, J., Blais-Morin, L.-A., Granger, E., 2021. Unsupervised Multi-Target Domain Adaptation Through Knowledge Distillation. In: Proc. IEEE Winter Conf. on Appl. of Comp. Vis, pp. 1339–1347.
Park, W., Kim, D., Lu, Y., Cho, M., 2019. Relational knowledge distillation. In: Proc. CVPR, pp. 3967–3976.
Redd, T.K., Campbell, J.P., Brown, J.M., Kim, S.J., Ostmo, S., Chan, R.V.P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J., 2019. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. Br. J. Ophthalmol. 103 (5), 580–584.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proc. CVPR, pp. 4510–4520.
Sen, P., Wu, W.-C., Chandra, P., Vinekar, A., Manchegowda, P.T., Bhende, P., 2020. Retinopathy of prematurity treatment: Asian perspectives. Eye 34 (4), 632–642.
Shen, Z., He, Z., Xue, X., 2019. Meal: multi-model ensemble via adversarial learning. Proc. AAAI 33, 4886–4893.
Son, W., Na, J., Hwang, W., 2021. Densely guided knowledge distillation using multiple teacher assistants. In: Proc. ICCV, pp. 9395–9404.
Tan, C., Liu, J., Zhang, X., 2021. Improving knowledge distillation via an expressive teacher. Knowl-Based Syst 218, 106837.
Taylor, S., Brown, J.M., Gupta, K., Campbell, J.P., Ostmo, S., Chan, R.P., Dy, J., Erdogmus, D., Ioannidis, S., Kim, S.J., 2019. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. JAMA ophthalmol 137 (9), 1022–1028.
Wang, J., Bao, W., Sun, L., Zhu, X., Cao, B., Philip, S.Y., 2019a. Private model compression via knowledge distillation. Proc. AAAI 33, 1190–1197.
Wang, J., Gou, L., Zhang, W., Yang, H., Shen, H.-.W., 2019b. Deepvid: deep visual interpretation and diagnosis for image classifiers via knowledge distillation. IEEE Trans. Vis. Comp. Graph. 25 (6), 2168–2180.
Wang, L., Yoon, K.-J., 2021. Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans. Pattern. Anal. Mach. Intell. 44 (6), 3048–3068.
Wang, S., Yu, L., Yang, X., Fu, C.-W., Heng, P.-A., 2019c. Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans. Med. Imag. 38 (11), 2485–2495.
Wang, X., Zhang, R., Sun, Y., Qi, J., 2018. KDGAN: knowledge Distillation with Generative Adversarial Networks. NeurIPS 783–794.
Xie, H., Lei, H., Zeng, X., He, Y., Chen, G., Elazab, A., Yue, G., Wang, J., Zhang, G., Lei, B., 2020. AMD-GAN: attention encoder and multi-branch structure based generative adversarial networks for fundus disease detection from scanning laser ophthalmoscopy images. Neural Netw 132, 477–490.
Yang, X., Lin, Y., Wang, Z., Li, X., Cheng, K.-T., 2019. Bi-modality medical image synthesis using semi-supervised sequential generative adversarial networks. J. Biomed. Health. Inform 24 (3), 855–865.
Yao, A., Sun, D., 2020. Knowledge transfer via dense cross-layer mutual-distillation. Proc. ECCV 294–311.
Zhang, H., Hu, Z., Qin, W., Xu, M., Wang, M., 2021. Adversarial co-distillation learning for image recognition. Pattern Recognit 111, 107659.
Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proc. CVPR, pp. 6848–6856.
Zhao, H., Sun, X., Dong, J., Chen, C., Dong, Z., 2020. Highlight every step: knowledge distillation via collaborative teaching. IEEE Trans. Cybern. https://doi.org/10.1109/TCYB.2020.3007506.
Zhao, R., Chen, X., Chen, Z., Li, S., 2021. Diagnosing glaucoma on imbalanced data with self-ensemble dual-curriculum learning. Med. Image Anal. 75, 102295.